The National University of Ireland, Maynooth

Developing the Implicit Relational Assessment Procedure (IRAP) as a Measure of Cheating Behaviour

Thesis submitted to the Department of Psychology, Faculty of Science, in fulfilment of the requirements for the degree of Doctor of Philosophy, National University of Ireland, Maynooth

Luis Manuel Silva

August 2015

Head of Department: Dr. Andrew Coogan

Research Supervisor: Professor Dermot Barnes-Holmes

Table of Contents

Abstract

The current thesis set out to investigate the suitability of the IRAP to assess attitudes in the moral domain and to predict cheating behaviour in a controlled context. Across six studies in two countries, we developed three IRAPs that targeted relations between actions and concepts of morality, reports of frequency of moral and immoral behaviour, and personal feelings towards engaging in moral or immoral actions, and interpreted our findings through the Relational Elaboration and Coherence (REC) Model. In the first part of the current research programme, correlations between the IRAPs and a cheating task suggested that individuals who are highly practised at immoral behaviour such as cheating, deceiving and lying are more likely to confirm that they do not engage in such behaviours, in itself an instance of that behavioural class. Further studies revealed that a history of bad feelings associated with engagement in immoral behaviour correlated with lower cheating, and that higher pro-moral biases in the IRAP correlated with lower reported psychopathic traits. In the latter part of the research programme described in the current thesis, a values-oriented intervention was shown to have an effect on IRAP performance and to produce a non-significant trend toward decreasing cheating levels. To conclude, strengths, limitations and opportunities for further research are discussed.

Introduction

The study of moral behaviour has a long history within Psychology, despite the difficulties of identifying its area of interest and scope with precision. The quality of "being moral" is in fact not easily defined: philosophers and scientists who have engaged in the study of moral behaviour, or "morality" as a dimension of behaviour, have advocated nearly every position ranging from the existence of moral universals shared by all human beings irrespective of culture and time, to moral systems entirely built upon social whim. For the most part, modern theories seem to establish a compromise between both extremes: a culturally mediated set of moral general, universal rules.

The word "moral" is defined as pertaining to the quality of being good or bad, both individually and socially, and it comes from the Latin term "mos", meaning "custom" (Hayes, Gifford & Hayes, 1998, p. 253). According to the Oxford Dictionary of English Etymology (Onions, Friedrichsen, & Burchfield, 1978), the word was first used by Cicero as a translation of the Ancient Greek word "ēthikós"; this is also the root for our word "ethics", which refers in turn to the study of morals and moral choices (Hayes, Adams & Rydeen, 1994).

In Western Philosophy, however, the two words have evolved to indicate two related but different things: the word "moral" describes both a set of prevailing behavioural guidelines within a culture and the ways in which the behaviour of individuals or groups adheres to (or departs from) these guidelines. The word "ethics" refers to the study what is moral, and is mainly concerned with offering behavioural guidelines based on philosophical and sociological reflection – the distinction is not unlike the one between theory (ethics) and practice (moral) (Ardila, 2014). However, the distinction between moral and ethics is not always observed and some authors have chosen to use both concepts interchangeably (Jones, 1991), since it is generally perceived that a component of "ought to be" is indeed present in

the realm of morality, which makes ethics redundant in non-philosophical settings. This is the approach we will follow in this document.

Another way of distinguishing between moral, ethics, and convention is to place them in a continuum marked by perceived importance of following a certain guideline, or the severity of its transgression. In this perspective, certain issues are ordinarily perceived to be moral in nature, such as the death penalty, incest, or abortion. Others, such as dishonesty and violations of professional codes seem to be within the realm of ethics, and disregard for local tradition or socially constructed rules of behaviour are assigned to the domain of convention. These are, however, very general guidelines and it is easy to find examples of situations that involve moral, ethical and conventional dilemmas at the same time.

Even though questions of morality and ethics have been historically dealt with by philosophers and, to some extent, politicians and legal professionals, they always refer to behaviours in context. Due to the pervasiveness of moral issues in human behaviour, it was only natural that social sciences tried to tackle the subject early on. The social scientific approach enriches the concept of morality by integrating the perceptions and beliefs of people in different cultures, which helps paint a more detailed picture of the separations and overlaps amongst the related domains of morality, ethics and convention. Some of this research, for example, suggests that the moral domain is perceived to be backed by a kind of prescriptive force independent of the power of authorities (i.e. divinities), and that transgressions of this domain are regarded as more serious than violations of convention (Kelly, Stich, Haley, Eng, & Fessler, 2007).

However, other research suggests that it is almost impossible to find examples of behaviours which could be universally considered immoral independently of culture, history and geography. For example, two cultures may differ in their consideration of funeral rites. Are they a moral obligation, or just a convention? For certain groups, lack of proper burial

may be as serious as incest, and for others it may be no worse than white lies. Moreover, some studies suggest that the distinction between moral and convention may be inherently flawed due to the nature of the tasks used to assess them (Kelly et al., 2007).

What seems to be clear is that people in a community, given appropriate context, can readily label certain behaviours as "good" or "bad". In a very general sense, these behaviours comprise the domain of "moral behaviour". Whenever such behaviours occur, a "moral issue" arises - a situation in which a person's behaviour brings either benefit or harm (Jones, 1991), and the individuals responsible for the emission of those behaviours, or affected by them, are called "moral agents". Traditionally, Moral Psychology studies the factors that influence decisions made by moral agents in situations involving moral issues.

Studying these moral issues is not always easy or straightforward, for two main reasons. First, moral judgment seems to depend heavily on contextual factors and on individual histories, which are difficult to cover completely using traditional measures. And second, people naturally tend to present themselves in a good light, even if it involves exaggerating their morality or, more often, underreporting their immorality. Therefore, explicit measures such as questionnaires or interviews entail the risk of capturing distorted or biased responses. Recently, however, researchers in the field of implicit cognition have been used special measures to assess socially sensitive topics such as prejudice and stereotype. These measures, as we will discuss later, seemingly capture attitudes that people explicitly conceal or are unaware of.

In the current thesis we present a programme of research on moral behaviour that uses one such measure of implicit cognition, the Implicit Relational Assessment Procedure (IRAP), to explore the moral domain and predict cheating behaviour. Our first port of call is a review of psychological theories of morality, which will provide context to the research proper.

Chapter 1

Theoretical Approaches to Morality

Theoretical Approaches to Morality

Until the consolidation of social sciences and cultural studies in the 20th century, it was philosophers who mostly had the monopoly of conceptual work on ethics and morals. However, social and psychological research has expanded continuously into the realm of morality, in an attempt to determine the factors that influence moral decisions and to create interventions that help decrease the frequency of unethical behaviour. Most accounts of moral behaviour in Psychology are influenced by cognitive science, and the core assumption is that moral behaviour (like any other) is the result of a series of internal cognitive processes. The most significant of those accounts are reviewed next.

Kohlberg's Moral Development Theory and its Revisions

Perhaps the most commonly used framework to study moral behaviour is the theory of justice reasoning (Levine, Kohlberg, & Hewer, 1985), a refined version of Kohlberg's previous cognitive moral development theory (Kohlberg & Hersh, 1977). The latter was inspired by Piaget's cognitive structuralism, and stated generally that cognitive structures responsible for moral reasoning gradually develop in a universal, culture-independent sequence. Both the original and the reformulated versions rely on an interviewing methodology which involves presenting individuals with moral dilemmas in the form of hypothetical situations, and requires them to answer a number of questions about the behaviours and motivations of the characters involved in the stories.

According to the theory, this method makes it possible to assess the level of development of the individual's justice reasoning processes. Broadly speaking, if the person's responses show that he or she tells right from wrong depending on the consequences of the behaviour (reward or punishment), their moral reasoning is said to be at a "preconventional" level. If, on the other hand, the action being right or wrong depends on

what society, community or authority agree upon, moral reasoning is said to be at a "conventional" level. But if the person comes to reason that right or wrong depend on reflectively constructed moral standards which are independent of the consequences to the self or others, his or her moral reasoning will be at a "postconventional" level.

Despite being widely known and having remained virtually unchallenged for some time, Kohlberg's theory of moral development has been criticised on at least four grounds: its emphasis on moral reasoning, the methodology used to assess it, the idea of sequential stages of moral development, and the universality of those stages across cultures (Burman, 1999; Shweder, Mahapatra, & Miller, 1987).

The first criticism states that if moral reasoning were the most important cause of moral behaviour, people who scored higher in moral development would have more refined moral reasoning processes, and their behaviour would therefore be more morally appropriate; however, research suggests that moral reasoning is not always related to actual behaviour (FeldmanHall et al., 2012; Shweder et al., 1987). In fact, even though individuals frequently report positive perceptions of their own moral behaviour, with most people describing themselves as kind, honest, compassionate, righteous, and caring (Aquino, Reed, Thau, & Freeman, 2007; Aquino & Reed, 2002), they also have trouble predicting and remembering unethical behaviour on their part (Tenbrunsel, Diekmann, Wade-Benzoni, & Bazerman, 2010). Both daily experience and controlled research reveal that a sizeable part of the population engages in behaviours that break their referential moral standards; these include corruption, cheating, and stealing, amongst others. As a matter of fact, such behaviours are practically endemic, in that people not only exhibit them rather frequently, but also seem to ignore others' immorality under certain circumstances (Gino & Bazerman, 2009).

This gap between moral judgment and behaviour challenges the idea that moral reasoning is an immediate cause of moral behaviour (Frimer & Walker, 2008). For Shweder et al. (1987), the situation is analogous to the fact that native speakers of a language can use grammatical decision rules properly and identify grammar mistakes, but they are not necessarily able to describe those rules. In the same way, people may not be able to appropriately describe their moral reasoning processes, but that does not mean they cannot use them to actually decide whether something is morally sound or not. In this line of thought, moral reasoning and its categories become just labels for a certain set of responses in a test, but are not really precursors of behaviour in other contexts.

Even if there were no such gap between behaviour and explicit reports of moral reasoning processes, the success of the interviewing methodology depends on the participant having relatively high verbal skills, including the ability to properly discuss complex and abstract ideas. However, this reliance on verbal argumentation can be affected by the fact that the possibility of knowing and using concepts and ideas is not necessarily correlated with the ability to discuss them accurately in speech. Starting with the seminal experiments by Nisbett and Wilson (1977), many researchers in the area of cognitive psychology have confirmed that participants can be aware of the results of their decision-making processes but rarely are they able to give accurate verbal reports of those processes themselves. More recently, Johansson et al. (2005) used a paradigm involving deception to show that people justify decisions that they in fact have not made. In short, over-reliance on verbal reports might produce skewed results due to the numerous factors that have an influence over them.

The idea of stages of moral reasoning is also a problematic one, since scientific evidence does not support the idea that cognition presents itself in clearly separated stages. In fact, research in the area of moral reasoning has shown that typical adults and children

tend to mix concepts and principles from different substages, and that it is very rare to observe the stages in pure form described by the theory (Shweder et al., 1987).

Finally, the universality of moral stages has been also challenged. Evidence in favour of this idea (for example Snarey, 1985) commonly consists of studies performed with populations much like the ones studied by Kohlberg and his colleagues (typically western, white, and urban). However, recent cross-cultural research on the subject has revealed, for example, that spontaneous descriptions of the moral domain made by people in the lower social classes or with religious backgrounds other than Christianity include factors such as duty, purity of mind, traditions, religious norms and others that are not part of the original model (Graham et al., 2011).

As a matter of fact, Kohlberg himself observed that some populations tend to score higher than others. This is usually explained by stating that processes of rational reasoning are unequally distributed across populations, although it may also be explained by a bias towards westernised elites in the theory (Shweder et al., 1987). Counter-intuitive findings, such as Tibetan Buddhist monks scoring lower in moral reasoning than ordinary populations (Gielen, 1983, cited by Snarey, 1985), also support this criticism.

The refined version of Kohlberg's theory of moral development addresses some criticisms by introducing a few changes: specifically, it states that it is concerned with "justice reasoning" (instead of moral reasoning), and relaxes the rigid stage structure of the original proposal. However, the methodology, the concept of stages, and the notion of fundamental, culturally-independent principles survive basically untouched (Levine et al., 1985), and the general idea remains intact, which extends most of the previous criticisms to this new version as well.

Post-Kohlberg Cognitive Theories

For the previous reasons, contemporary moral Psychology has started to explore other factors beyond moral reasoning, and cognitive approaches have flourished. Perhaps the most well-known criticism of Kohlberg's theory was raised by Carol Gilligan (1982), who proposed that men and women have different moral orientations and moral developmental pathways related to their gender-specific traits, and that most accounts of morality are based on the typically masculine orientation towards justice (as opposed to care in women). Her theory involves substantial changes not only to the conceptual basis of morality that had been stated up to that point, but also to the methods used to assess moral development (Walker, 2006).

Other, more general approaches have focused on social and cultural factors beyond gender. One of the most recognised is Social Cognitive Theory (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996), which states that people develop moral standards and will usually do things according to these moral standards, due to the existence of self-monitoring and self-regulatory cognitive processes. However, there are many different mechanisms through which people can temporarily soften or relax their moral standards, and thus engage in actions that go against them ("moral disengagement"). These mechanisms are used because the misalignment of actions and goals creates cognitive dissonance, a psychological state of discomfort. The reduction strategy is either a change in behaviour, or a temporary change in morals (Shu, Gino, & Bazerman, 2011).

According to the theory, the mechanisms involved may operate during the various steps of the self-regulatory process. The reprehensible behaviour may be justified, euphemistically labelled, or compared to other situations as a means of reducing its impact; the effects of this behaviour may be minimised, ignored or misconstrued, and the victim may

be dehumanised or made responsible for the negative consequences. And if those are not enough, displacement or diffusion of responsibility can also happen (Bandura, 2002).

Other cognitive approaches have also been described. In general, most propose that moral behaviour is the product of a four-staged process: a) awareness of an ethical issue, b) ethical judgment, c) establishment of an intention to behave ethically, d) and actual display of the behaviour (Reynolds, 2006). However, cognitive perspectives have been challenged by theories drawing from evolutionary science and neuroscience. One of the most recent approaches, which illustrates the influence of evolutionary theory, is Moral Foundations Theory (MFT; Haidt & Joseph, 2004). It states that humans are genetically endowed with the possibility of developing concern for a small number of moral intuitions related mainly to protection of kin, reciprocity, group cooperation, respect for authority, and avoidance of microbes and parasites (Graham, Haidt, & Nosek, 2009). MFT thus differs from previous theories in that moral reasoning mostly happens after innate moral intuitions have played their part. However, empirical evidence is still scant (Graham et al., 2009), and some criticism has also appeared based on MFT's ideas of innateness and modularity, and that there may be other strong candidates for additional moral foundations (such as industriousness) (Suhler & Churchland, 2011).

The idea of moral reasoning not being the sole precursor of moral action is not unique to MFT. Intuitionist models of morality have suggested that moral reasoning may simply be a way of justifying the moral judgment or the moral behaviour after it has already been performed, in a sort of rationalisation of moral intuitions (Haidt & Joseph, 2004). An emphasis on intuitionism suggests that whatever mechanisms lead to moral behaviour are of a more unconscious, irrational nature, probably related to the evolutionary history of the species.

A Discussion of Cognitive Views of Morality in Psychology

The dominant paradigms in Moral Psychology stem from cognitive perspectives. These perspectives regard observable behaviour as the result of the operation of internal structures or processes on external and internal stimuli. The nature and features of those internal components are inferred through the use of different assessment methods, ranging from the simple and mundane, like behavioural observation and interviews, to the highly sophisticated, such as electrophysiological measures and neuroimaging.

A problem with this view is that the mental or cognitive structures and processes that are said to cause behaviour cannot be observed directly: their existence and mode of operation can only be inferred with reference to the conditions in which they presumably operate (observable features of a context or situation), the biological events that are assumed to underlie the mental events themselves, or to their products (actions). This has resulted in several different (and sometimes conflicting) models of cognitive systems, and also, as previously described, in contradictory or counterintuitive findings that limit research and conclusions derived from it.

One example of these counterintuitive findings is the problem of the discrepancy between verbal descriptions that people give of their moral reasoning processes and their actual moral behaviour, which presumably results from those processes. For instance, a person may overtly say that he or she believes that no one should be punished to death, yet when called for jury duty may act by voting for capital punishment of an individual. If moral actions are guided by internal moral processes or beliefs, little or no discrepancy should be found. However, it would seem that people think and act in a certain way, but are able to report that they think and act differently.

Even though cognitive perspectives do acknowledge that social processes are conditions that shape moral judgment, the main explanatory mechanism for behaviour generally remains the action of internal and individual factors that account for the presentation of behaviour. This view is so pervasive that it has hardly been questioned at all, even in modern scientific literature. For example, a relatively recent review of morality considers it to be "a *mental phenomenon* that consists in thoughts and feelings about rights and duties, good and bad character traits (virtues and vices), and right and wrong" (Krebs, 2008, p. 150, emphasis added).

However, the shortcomings of these views have led psychologists and other social scientists to seek other assessment methods and explanations, specifically exploring whether genetics or physiology could contribute to a more complete account of morality, either as innate processes or as biologically mediated experiences. This biological focus was considered very briefly above, but a more detailed summary of the contributions of research on the biological factors participating in morality is presented in the next section.

Biological Factors linked to Moral Behaviour

Ethological views

Ethologists and biologists have generally maintained that considering human morality an exceptional case in nature does not follow the central tenets of evolutionary science, where morality is part of human nature and is also the result of evolutionary processes that have been operating for thousands of years. According to a recent view, interdisciplinary studies suggest the presence of moral building blocks, automatic moral judgments and intercultural similarities in moral domains such as fairness, reciprocity and empathy, and taken together this provides evidence for a biologically-determined layer of moral decision-making (De Waal, Smith Churchland, Pievani, & Parmigiani, 2014)

Studies with animals have observed behaviours in non-human species that, if emitted by humans, would raise moral issues. Indeed, one can speculate that the presence of such behaviours in different species makes the case for phylogenetic inclinations to immoral behaviour in humans, which are modulated and modified by social and cultural influences upon the development of language. We will discuss this possibility later in more detail from the psychological perspective of Relational Frame Theory.

For the time being, it seems important to note that through the examination of "immoral" behaviours in other species, evolutionary science has provided ethological perspectives to the study of morality. The most obvious example of *deception*, a moral issue if applied to human affairs, is mimicry, widespread in the animal and plant kingdoms, but not really behavioural in nature since it normally involves camouflage not under the control of the organism (for example, eye-shaped patterns on butterfly wings).

However, certain animal behaviours have been observed that are akin to what is called "immoral" or "unethical" in humans. Primate deception, for example, occurs both actively (e.g., a false anti-predator call in order to take advantage of the momentary distraction) and passively (e.g., withholding information about a food source), and is mainly related to the decreased availability of food (Wheeler, 2008). For example, capuchin monkeys seem to use deceiving alarm calls more frequently when both the amount and location of the available food makes it more contestable, and when the individuals themselves are in a spatial location that maximises their feeding success if their peers answer to the deceptive call (Wheeler, 2009). However, deception is not only related to the availability of food resources: gelada baboons, for instance, exhibit increased likelihood of extra-pair copulation, and less accompanying vocalisation, when the cuckolded male is a large distance away (Le Roux et al., 2013).

Perhaps it is unsurprising that other primates present human-like unethical behaviour, given our similarity to them and the ubiquity of deceptive behaviours in humans, but there are also examples in lower-order species. The giant cuttlefish, a colour-changing cephalopod, can fool other males by displaying female patterns on the skin on one side and male patterns on the side visible to females (Brown, Garwood & Williamson, 2012). The dance fly (Rhamphomyia sulcata) and spiders of the Lycosoidea family are some of the species in which males present "gifts" (prey) to females in order to increase their mating chances, and deception has been observed in both species with regards to those gifts: Lycosoidea male spiders wrap their gifts in silk, but occasionally "reuse" gifts that were rejected by the females or simply give them empty silk packages, which results in nearly equally increased mating opportunities. Dance fly males also give inadequate or false gifts on occasion, deceiving the female (Albo, Winther, Tuni, Toft, & Bilde, 2011).

Even though these behaviours are normally emitted in response to certain environmental conditions (a decreased availability of resources being the main example), the fact that their properties change on occasion suggests that more elaborate behavioural and perceptual processes than reflexes or fixed action patterns are involved in their presentation. Even though they do not reach the levels of complexity involved in decision-making in verbally sophisticated humans, such actions may not be completely unrelated to human morality or at least the evolutionary basis of moral/immoral behaviour in humans.

Other factors related to biology

Biologically oriented perspectives have also addressed other biological factors that may be related to moral behaviour. One of them is the depletion of cognitive resources due to tiredness; for example, Barnes, Schaubroeck, Huth and Ghumann (2011) found reduced self-control on a cognitive task in participants reporting fewer hours of sleep, and also more unethical behaviour correlated with less sleep.

Self-regulatory resources seem to decrease throughout the day, which could explain the so-called "Morning Morality Effect" (Kouchaki & Smith, 2013) – the presumed higher prevalence of immoral behaviour during the afternoon hours. Cognitive depletion can also result from tasks that involve response inhibition, and again, it seems to have an effect on subsequent responding in situations related to moral issues (Muraven, Pogarsky, & Shmueli, 2006).

While the role of fatigue on moral judgment has only been explored recently, an older candidate is physical disgust. People have long been using expressions containing words related to disgust to describe moral transgressions, their perpetrators, and their own feelings with regards to them. Some studies have found correlations between the strength of moral judgments and sensitivity to physical disgust (Chapman & Anderson, 2014; A. Jones & Fitness, 2008), although a recent study seems to have detected a dissociation between moral judgments and the elicitation of disgust when using electrophysiological measures (Yang, Li, Xiao, Zhang, & Tian, 2014) . However, others believe that people use disgust-related words to describe events in the moral domain out of convenience, but they are really invoking the basic emotion of disgust anyways (Nabi, 2002).

Having presented the main theoretical approaches and reviewed some of the biological and cognitive factors that influence moral behaviour, it is time to turn our attention to the general methodological approaches to the assessment of moral behaviour and the issues faced when employing them.

General Methods for the Assessment of Moral Behaviour

As mentioned earlier, each theoretical approach seems to have a preference for a certain subset of methods. Classical Moral Psychology from the cognitive perspective relies on interviewing methods and standardised psychological testing, sometimes including

projective techniques which have little empirical support (Lilienfeld, Wood, & Garb, 2000). More contemporary approaches advocate the use of physiological measures such as skin conductance, electroencephalography and magnetic resonance imaging. We will now discuss the advantages and disadvantages of these methods of assessing moral behaviour.

Self-report methods. A simple way of assessing moral behaviour is to use interviews and self-reports. These are easy to use and require little more than pen, paper and a desk. Kohlbergian approaches, for example, use a form of interview in which verbal responses to imagined situations ("moral dilemmas") are used as a device to assess the state of an individual's moral reasoning processes. An example of one such dilemma tells the story of a man considering stealing a drug that might save his wife's life after not being able to come up with enough money to pay for it and finding his pleas to the seller rejected. The participant is asked a series of questions such as "Should he steal the drug?" and "Is it right or wrong to steal it".

The moral dilemma scenario has been used in other ways to study how the presentation of the situation impacts the answer given. A classic example is the "trolley problem" (Klein, 2011), which has two variations that generally produce different results, although the problem is essentially the same. In the first variation, people are asked whether they would flip a switch to change the course of a runaway train carriage (i.e. trolley), and by doing so getting it to kill a bystander, but saving five others instead. The second variation asks whether individuals would push the bystander onto the track so as to stop the trolley and save the other five. Far more people are willing to flip the switch than to push the bystander, even though the result is the same (the person interviewed would save five lives at the expense of one).

The main issue with self-reports to study socially sensitive topics, including moral behaviour, is that they have been long proven to be highly susceptible to cognitive biases and

distortions. For instance, research on personality assessment through questionnaires and scales has shown that responses vary in the presence of certain contextual cues and that the overall results can be faked by respondents trying to present a certain image of themselves (Holden, 2007; Krahé, Becker, & Zöllter, 2008).

Studies involving other situations where participants are requested to report their performance frequently find discrepancies between reported and actual performance, and examples are numerous: reporting donating to charity and not doing so (Bekkers & Wiepking, 2010), or declaring having been offered gifts or services in exchange for votes only when asked in an anonymous survey but not when questioned face-to-face (Gonzalez-Ocantos, de Jonge, Meléndez, Osorio, & Nickerson, 2012), to name a few.

In fact, as early as the 1950s, psychologists had already noticed the inaccuracies of introspection and observed that responses to clinical measures were commonly affected by a tendency towards socially desirable responding, which prompted the development of scales to assess this type of bias in both clinical and non-clinical populations (see for example Crowne & Marlowe, 1960). Many different measures and studies tried to isolate social desirability during the following decades, and despite some research suggesting that its role was being exaggerated, there appears to be evidence for a sizable enough effect, which demands a need to control for this variable in order to reach valid conclusions (Holden, 2007).

To account for these inaccuracies, researchers have hypothesised that giving overt answers to questions, either orally or in a questionnaire, involves a controlled (i.e. conscious) process in which respondents have enough time to analyse the question and become aware of the social implications of their answer. The final response will, to some degree, be affected by the results of this process, and thus may not accurately reflect "true" or actual personal beliefs .

Whatever the possible explanation may be, it is clear that verbal reports of mental activity seem to be strongly susceptible to the influence of cognitive bias due to limited access to said activity or, more commonly, to the perceived need for adapting reports to social convention (Hughes, Barnes-Holmes, & De Houwer, 2011). Some researchers have tried to circumvent the problems of accessibility and self-presentation by using modern neuroimaging technology to try to find the components of moral decision making in the brain.

Neuroimaging. In general, the use of functional magnetic resonance or other real-time brain imaging techniques suggest that immoral behaviour involves the intentional suppression of a default truth-telling (moral) response (Verschuere, Spruyt, Meijer, & Otgaar, 2011), which seems to include the participation of several areas of the brain. For instance, a meta-analysis by Christ, Van Essen, Watson, Brubaker and McDermott (2009) supports the critical role of prefrontal areas in deception, most likely due to their participation in executive control, which is regarded as an important component in producing deceptive responses; for example, participants asked to lie show increased activity in the bilateral ventrolateral prefrontal area (Spence et al., 2001). The ventromedial pre-frontal cortex has been suggested to be important in the perception of harmful intent, which in turn modifies moral judgment (Young et al., 2010).

Other areas whose role in deception has been studied include the posterior superior temporal sulcus and the amygdala (Stanley, Phelps, & Banaji, 2008), which seem to respond differently when evaluating positive and negative deviances from moral standards (Takahashi et al., 2008). Further research has suggested that the participation of mesencephalic and diencephalic structures can account for the non-conscious and intuitive components of moral decision-making (Reynolds, 2006).

Findings from neuroimaging studies are generally used as supporting evidence in the creation of cognitive models that intend to account for the hypothesised internal processing that results in the emission of behaviour. For example, the different responses in the two versions of the aforementioned "trolley problem" are accounted for from a cognitive perspective by theorising that morality works via a dual-process system, in which the wording of the problem and the nature of the situation primarily engages either the cognitive or emotional components of moral judgment. This roughly corresponds with neuroimaging findings of increased posterior cingulate and superior temporal activity during consideration of emotional dilemmas ("push and kill someone to save many") in which participants are asked to imagine causing direct harm, and increased activity in the inferior parietal lobes and the dorsolateral prefrontal cortex during reasoned judgments ("one must die for the greater good") (Klein, 2011).

There are two main problems with assessing moral behaviour from neuroscientific perspectives. The first one has to do with resources: experimental paradigms are complex and require resources and technology, such as functional magnetic resonance imaging (fMRI) devices, that are not necessarily readily available to most researchers due to their cost and their complexity. For this reason, researchers have sought to devise cheaper, simpler and less invasive alternatives within the experimental context, in order to control for interpretations and biases.

The second and more critical problem with neuroimaging studies of immoral behaviour is that despite the evidence for the participation of certain brain areas in moral responding, studies have been unable to identify systematic differences in activation patterns that enable researchers to separate moral from immoral responses (Verschuere et al., 2011). Moreover, the areas identified have been also shown to have important roles in processes

such as working memory, task switching and inhibitory control, making it difficult to separate the specific components of moral behaviour (Christ et al., 2009).

Electroencephalography and involuntary motor actions. The aforementioned difficulties in the use of neuroimaging have led some researchers to turn to less resource-intensive techniques which sacrifice spatial resolution but increase temporal resolution, such as standard electroencephalography (EEG), event-related potentials (ERP) and regional cerebral blood flow (rCBF). Of these, ERP, the analysis of brain activity related to the presentation of a stimulus, is the best candidate for a general method of assessing moral/immoral behaviour. Ortu (2012) suggested, for example, that the P300 (a positive electrical peak in the EEG signal approximately 300ms following the onset of a stimulus)could be used as a marker of deception. Using deception and concealment of information as operationalisations of unethical behaviour, several studies have found particular ERP patterns in tasks involving detection of simulated amnesia (Rosenfeld et al., 1998; Rosenfeld, Ellwanger, & Sweet, 1995), deception (Johnson & Rosenfeld, 1992; Spence & Kaylor-Hughes, 2008) and concealed information in mock crime experimental paradigms (Rosenfeld et al., 2008). Nevertheless, it has been argued that the P300 (and related ERP patterns) could merely be a reflection of an orientation response or a shift in attention and is not directly related to immoral behaviour *per se.*

On balance, direct measurements of cerebral activity may not be the only physiological indicator of moral behaviour. Several experimental findings suggest that certain motor behaviours occur before conscious awareness of a moral judgment or correlate with psychometric measures but not with self-reports about certain tasks. This idea is known as the "ideomotor principle", which presumes that ideas or thoughts can occur together with involuntary motor actions (Stock & Stock, 2004), and from the observation of these actions, the corresponding cognitive activity could be inferred. A key component of the ideomotor

principle is that those involuntary motor behaviours are not necessarily accompanied by conscious awareness, which suggests that their presentation is less susceptible to voluntary distortion. At the time of writing, however, systematic evidence to support the use of the ideomotor principle to examine moral behaviour was unavailable.

Response latencies. Some of the studies described before have found a correlation between ERPs and response latencies (Rosenfeld et al., 1998; Seymour, Seifert, Shafto, & Mosmann, 2000), which suggests that removing the electrophysiological component and only using latencies can still be useful in the assessment of moral behaviour. The use of response latencies –the elapsed time between the presentation of a stimulus and the emission of a particular response – has a long history in psychology, because it was one of the favourite measures used in psychophysics, during the first decades of the development of scientific psychology.

In the moral domain, response latencies have been used in experimental paradigms where moral issues are operationalised by putting people in situations in which they can conceal information or tell lies. Such actions have been suggested in a number of studies to involve a greater degree of cognitive control, which in turn results in increased reaction times when participants take these tests. For instance, Spence et al. (2001) interviewed participants to get a baseline of simple actions that they had recently performed and were then instructed to lie on some of them in the presence of an observer who would ostensibly try to tell which responses were true or false. Upon comparing both sets of answers, significantly different response times during lies and truths were found.

More recently, Noordraven and Verschuere (2013) used a mock crime scene paradigm to assign participants to a guilty group, who had advance knowledge of a crime that was supposed to take place, and an innocent group without such knowledge. They found significant differences in reaction times between both groups using the Concealed

Information Test (CIT), with the guilty group having higher response latencies than the innocent group. Similar results were obtained by Williams, Bott, Patrick and Lewis (2013), who tried to control for extraneous factors by asking participants to lie or tell the truth about the shape of a figure presented on the screen, and found that, on average, telling lies took slightly more time. In general, the most widely supported explanation for this effect, as previously mentioned here, is that telling lies entails suppression of a default truth-telling response, and the extra cognitive workload involved explains the difference in response times (Vendemia, Buzan, & Green, 2005).

Of course, a reaction time measure would only be useful to assess moral behaviour if it successfully resisted the individual's attempts to change it or fake it – that is, if it were an accurate, non-changeable somatic marker of telling lies or behaving immorally (Sobhani & Bechara, 2011). Only a few studies have so far concerned themselves with investigating this possibility, and the evidence seems to be inconclusive. On the one hand, for example, Vendemia et al. (2005) found that practice does not seem to have a significant effect on reaction times. In a more recent study, however, latencies associated with lying decreased when participants often told lies during an earlier part of the experiment, and conversely, lying became more difficult (i.e., increased latencies) after frequent truth-telling. (Van Bockstaele et al., 2012).

Implicit measures. Response latencies have also been used within a booming domain in Experimental Social Psychology: implicit testing. Researchers in the field of implicit cognition suggest that there are automatic, meaningful responses which can be modified by subsequent conscious, socially-mediated assessments of the potential consequences of those responses. Essentially, the adjective *implicit* used in this manner is synonymous with *automatic* and describes psychological processes that require few

cognitive resources, are relatively resistant to change and are present even in the absence of awareness and particular goals (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009).

Implicit testing borrows from the so-called response priming paradigm, which uses response latencies to study issues of early information processing. A typical response priming experiment involves responding (e.g. pressing a key) quickly to a target stimulus which is preceded by a prime stimulus. The prime can be mapped to either the same (consistent prime) or the opposite (inconsistent prime) response to the target stimulus, and the generally observed effect is that consistent primes result in faster responses to the target, whereas inconsistent primes typically slow the responses (Schmidt, Haberkamp, & Schmidt, 2011).

A sample trial presents a sample shape (diamond or square) for a few milliseconds, and then a target shape which would either be the same as the prime (consistent) or a different one (inconsistent). Two different response keys are assigned to each target, and it is predicted that response latencies are lower in consistent responding (i.e., both prime and target are the same) – this is the priming effect.

A variation called affective priming involves the use of emotionally-loaded stimuli (such as pictures of faces or situations, or words describing emotions), and the differences in reaction times when responding to targets after consistent and inconsistent primes are hypothesised to reflect attitudes towards the affective primes (De Houwer et al., 2009).

Evidence for response and affective priming comes mainly from two sources: the implicit cognition literature and vision research; the latter because the procedure has been used to study observed relationships between motor control and visual awareness (Schmidt et al., 2011). In their meta-analysis, De Houwer et al. (2009) state that studies have established that affective priming tasks are a useful device to capture attitudes, despite the

well-known fact that priming effects may be produced by features of the stimuli other than the evaluative dimension (e.g. target depletion, De Houwer, Hermans, & Spruyt, 2001; signal strength, Francken, Gaal, & de Lange, 2011; previous learning, Horner & Henson, 2008).

A number of tests are currently available for the assessment of implicit cognition, most of them based upon the paradigm of rapid responding in computerised categorisation tasks, with reaction time and response accuracy as indicators of the implicit attitude. The stimuli involved in the categorisation tasks are manipulated according to the topic towards which the implicit attitude is exhibited. We will now present a summary of those available implicit measures.

The Implicit Association Test and related measures. According to Greenwald, McGhee and Schwartz (1998), implicit attitudes, which reflect automatic, non-conscious evaluation, can be tapped into if participants are asked to categorise stimuli both as accurately and quickly as possible. This is because latencies are a function of the degree to which concepts are associated in memory, with lower latencies meaning that two stimuli are firmly related in a given individual's cognitive structure, and thus have the power to influence behaviour.

To test this idea, Greenwald, et al. (1998) designed the Implicit Association Test (IAT), which has become the best known implicit measure, and it is currently supported by extensive evidence: over 700 papers using the IAT have been published since its inception in the late 1990s. Studies using the IAT have dealt with such diverse topics as consumer preferences (e.g. Ayres, Conner, Prestwich, & Smith, 2012; Piqueras-Fiszman, Velasco, & Spence, 2012), cultural perceptions of body image (e.g. Brewis & Wutich, 2012), violence (e.g. Eckhardt, Samper, Suhr, & Holtzworth-Munroe, 2012), and, especially, racial, ethnic and gender stereotypes (Haider et al., 2011; Rezaei, 2011; Roddy, Stewart, & Barnes-Holmes, 2011; Rooth, 2010; van Ravenzwaaij, van der Maas, & Wagenmakers, 2011). Under the

weight of this accumulated evidence, there is now a general consensus that IAT scores reflect implicit attitudes, at least sometimes and to some degree (De Houwer et al., 2009).

The basic structure of the IAT and many of its derivatives involves asking participants to categorize two stimuli together by pressing one response key and to categorize another two stimuli together using a second response key. In a race-related IRAP, for example, one block of trials might involve pressing one key whenever a picture of a white person or a positively valence word is presented and pressing a second key whenever a picture of black person or a negatively valenced word is presented. In another block of trials the categorisation responses are reversed, such that one key is pressed for white pictures and negative words and the second key is pressed for black pictures and positive words. Typically, white participants find it easier (i.e. produce lower response latencies) when they have to respond White+Positive and Black+Negative compared to when they have to respond White+Negative and Black+Positive.

The mechanism by which this type of IAT effect is produced has been disputed. Greenwald et al. (1998) first suggested that the results of IAT tasks are related to the degree of association between two concepts in memory. During the following years, efforts were undertaken to elaborate on this idea and give a more detailed account of the process. Brendl, Markman and Messner (2001) have suggested that the effect is produced in a random-walk model in which both valence and identity of the stimuli compete or collaborate to produce a response depending on whether a consistent or inconsistent response is required. Other mechanisms, including response-activation effects and differential response costs derived from task switching have also been proposed (De Houwer, 2001).

Despite its widespread use and apparently high validity, some limitations have been identified with the IAT. It has been established that the nature of the stimuli used, the differences in cognitive abilities, and order effects may be partially responsible for IAT effects

(De Houwer et al., 2009). In order to address these and other limitations, other tests such as the Go/No Go Association Task (Nosek & Banaji, 2001), the Extrinsic Affective Simon Task (EAST, De Houwer, 2003) and the Brief IAT (Sriram & Greenwald, 2009) have been developed. The detailed methodological issues surrounding the development of these types of alternative tests are beyond the scope of the present chapter. The critical issue, for the current research revolves around the theoretical claim that the IAT and the vast majority of implicit tests are based either explicitly or implicitly on the notion that implicit cognition is inherently associationistic. That is, excitatory or inhibitory links between internal representations of stimuli are said to be passively formed under certain environmental and organismic conditions by pairing of the stimuli. These links enable activation of one of the stimulus through activation of the other and it is these activations that are captured by the IAT and related measures (Hughes et al., 2011).

Critically, however, recent studies have been providing some evidence for the predictions made by alternatives to associationist views, such as those posed by propositional and functional-contextualistic theoretical perspectives. For example, it has been found that non-evaluative propositions can influence automatic evaluative responses and that implicit attitudes might be formed not only by pairing stimuli, but through other sources of information such as instructions (De Houwer, 2006) or even interactions between parents and siblings (Castelli, Zogmaister, & Tomelleri, 2009). Other recent findings, such as the ability of people who show strong stereotypes to make intracategory differentiations (Scherer & Lambert, 2009) and the well-established fact that indirect procedures do not provide exclusive and unimpaired access to automatic processing, but also reflect some controlled processing at least, also suggest that there is more to implicit attitudes than association between concepts and evaluations (Hughes, Barnes-Holmes, & Vahey, 2012).

For these reasons, interest in pursuing alternative conceptual frameworks and improved assessment methods has rekindled. One such programme of research that has grown quite rapidly in recent years is the emergence of a behaviour-analytic or functional-contextual approach, driven largely by a modern behavioural account of human language and cognition, known as Relational Frame Theory (RFT; Hayes, Barnes-Holmes & Roche, 2001). This perspective provides the conceptual bedrock for the empirical research presented in the current thesis and thus we will consider it in some detail. Before doing so, however, it seems wise to explain how this relatively novel approach to implicit cognition emerged from the behaviour-analytic tradition.

Chapter 2

A Behavioural View of Morality

A Behavioural View of Morality

As we have seen, with the influence of neurocognitive and evolutionary perspectives, explanations of moral behaviour have drifted from strictly rational, individualised approaches, to more culturally-sensitive processes of moral decision-making with non-conscious, automatic components. Still, most psychological approaches to moral behaviour consider it to be a product of cognitive processes (moral reasoning and self-regulatory processes, both related to activity in certain areas of the brain) which are influenced by both genetic endowment (moral foundations) and social experiences.

More recent models, such as the neurocognitve approaches, also provide new methods of assessment. Even though studies employing physiological measures such as EEG, event-related potentials, and cerebral blood flow are to be found in the literature, the equipment needed and the difficulties for interpretation inherent to their use has prevented this type of research from taking a more prominent role. Thus studies of morality from the neurocognitive perspective remain a very small minority of the published literature.

In general, cognitive (including neurocognitive) theories and explanations attempt to explain moral behaviour by appealing to some form of mental mechanism or processes through which moral reasoning or judgments occur, which then serve to control moral or immoral actions. As such, moral decisions may be explained through different cognitive mechanisms, and the general purpose of cognitive research is to determine which mechanism or processes provide the best explanation. As we will see, the behavioural view contrasts with cognitive accounts in that it does not concern itself with hypothesised or inferred internal cognitive mechanisms as an explanation of behaviour and seeks to identify and analyse the functional relationships between behaviour and environmental conditions and events (Hayes et al., 2001).

Skinner's Operant Account and its Criticisms

According to Soreth (2011), a behaviour analytic account of morality is based on the rejection of internal agents as causes of moral behaviour, the recognition that human rights and morals are culturally dependent (anti-foundationalism), and the role of reinforcement as a sort of universal principle that governs behaviour. These principles are exemplified by Skinner's treatment of morality, a concise version of which is presented in *Beyond Freedom and Dignity* (Skinner, 1971).

In the book, Skinner propounds that science can provide not only answers to questions of possibilities (what people can do), but to questions of duties as well (what people ought to do). The latter are generally perceived to correspond to value judgments, out of the realm of science. A behavioural account, however, posits that labels such as good or bad, or right or wrong, can be applied to stimuli and behaviour, and this classification will essentially refer to their positive and negative reinforcement properties (Skinner, 1971, 1975).

A basic example would be tasty food. Eating it increases the probability that we will eat it in the future, and hence we can say it is "reinforcing". Of the food we say it is "good" or "delicious", which are verbal labels we assign to positively reinforcing things. Things that we label as bad are those that negatively reinforce us, such as physical pain: its disappearance reinforces the behaviour that enables us to avoid it. However, these reinforcing effects need not only be biological in nature, because a verbal community can also teach its members to value something as good or bad, and conditioned reinforcers (e.g., money) can also be held as such.

In a view of morality as the realm of what is good and bad, of what is and ought to be, the behavioural processes of positive and negative reinforcement are the foundations of what

other perspectives in psychology have called a “sense of morality”, meaning the attachment of value judgments to both environmental stimuli and behaviours. This sense of morality develops as children become more experienced with the “moral” labels shared by their verbal communities.

It is important to add that, in a behavioural perspective, there is no causal relationship between moral reasoning and moral behaviour, because any instance of moral reasoning is verbal behaviour potentially influenced by a different set of contingencies; that is, the environmental arrangements or conditions that make an instance of moral reasoning possible are probably not the same that create an instance of the actual moral behaviour.

For instance, a person may verbally state a moral judgment such as “I think terrorists deserve capital punishment and I would carry it out for my country and for freedom”. However, when the opportunity to act according to said moral judgment presents itself, the person might not be able to do the deed, and this is due to functional relationships between both behaviours and separate sets of contextual conditions. In the first case, the statement of the moral judgment is probably influenced by public outrage in the media, or beliefs presented by the reference cultural group, or a desire to look “strong” and “patriotic” in front of others. However, upon having to actually carry out an execution or otherwise act in such a way that the salience of causing a death is increased, other sets of influences become apparent: perhaps “thou shalt not kill” or “I do not want my children to remember me for this”.

The behaviouristic perspective has been traditionally considered by some professionals to be unable to account for complex human moral behaviour. The basic argument is that reinforcement as a determinant of behaviour could not possibly explain the subtleties of moral judgment. This particular criticism is raised when trying to account, for example, for the different responses to the trolley problem described above. That is, the

consequence in both cases is the same (i.e., 5 people survive because 1 person is killed), and yet participants respond differently depending on how the "moral dilemma" is presented to them. Thus an explanation for a moral judgement simply in terms of the reinforcing consequences that are arranged for a particular act does not seem to apply here. On balance, behavioural psychology recognises that reinforcement does not exist independently of other contextual arrangements, and factors such as discriminative stimuli, particular learning histories and verbal behaviour need to be taken into the explanation.

In the end, assessment of moral behaviour from the Behavioural perspective is based on one of the central tenets of Behavioural Psychology: that behaviour is in itself a legitimate object of study for Psychology, and is not to be regarded simply as the by-product of internal processes or states. This means that traditional measures of morality, such as those derived from Kohlberg's theory, are, at least, not to be interpreted in the same way from a behaviour-analytic point of view, since they presume that what is being assessed is the current state of a set of cognitive structures that are responsible for the appearance of certain behaviours. This is not to mean that they cannot be used at all, but only to point out that the information they provide is framed in a particular way, different from mainstream Psychology. For example, the very concept of "moral behaviour" in Behavioural Psychology involves verbal behaviour ("labelling") related to the behaviour of interest, because it has no inherent moral value – as said before, behaviour in itself is neither good nor bad.

Perhaps the main criticism of the traditional behavioural approach to the psychology of morality is that the preponderance of basic learning processes in explanations of behaviour fails to capture and clarify the role of language in human interaction and learning. Skinner recognised that language is a powerful modulator of experience, and some of his concepts have been employed widely and with considerable success in teaching basic language skills to learning disabled populations (Dymond, O'Hora, Whelan, & Donovan,

2006). However, his works on verbal behaviour failed to generate systematic and productive programmes of research.

This failure has been attributed to an inadequate definition of verbal behaviour as behaviour of the speaker that is modulated by the behaviour of a listener constitutes a departure from the functional definition of all other behaviours. Specifically, this definition is not made in terms of the learning history of the speaker (Hayes & Hayes, 1989; Hayes et al., 2001). For example, the behaviour of a rat inside an operant chamber may be mediated by the behaviour of the experimenter, who was trained by a verbal community on how to perform that mediation, and Skinner explicitly defines that operant behaviour of the rat as verbal. As pointed out by Hayes, et al, however, if the same rat in the same chamber obtains reinforcers "accidently" (e.g., lever presses knock food pellets into the chamber by nudging a torn sack of pellets resting against the side of the chamber), the rat's behaviour is rendered non-verbal. In effect, the distinction between verbal and non-verbal behaviour of a specific organism is not defined in terms of the behavioural history of that organism, but in terms of the behavioural history of a separate organism (i.e., the listener). This constitutes a clear departure from how other functional definitions are rendered in behaviour analysis.

New Directions in Behavioural Treatments of Verbal Behaviour

Attempts to devise productive programmes of basic research using the categories proposed in Skinner's treatise about verbal behaviour resulted mostly in cumbersome processes of data collection and analysis (e.g., the Reno Methodology) that yielded few insights, at least within the context of a basic research agenda. or research that so closely resembled traditional operant studies that the categories proposed in *Verbal Behavior* could be dispensed with (Hayes et al., 2001). It is worth noting, however, that Skinner's (1957) taxonomy did lead to some success in the area of applied behaviour analysis in terms of developing protocols for teaching specific language skills to learning disabled populations

(Sundberg, Partington, & J.W., 1998). Nevertheless, the field of basic research on human language and cognition within behaviour analysis stagnated for a few decades, and the problem of how to account for the florid nature of human verbal behaviour remained unsolved.

However, in the early 1970s the seminal work of Murray Sidman on the phenomenon of stimulus equivalence provided key elements for a better understanding of verbal behaviour. Sidman (1971) trained a young boy with learning disabilities to match a set of printed words he had not seen, with their corresponding spoken forms, which he could already match to drawings. After the training, it became apparent that the boy could also match the printed words to the figures, even though that relation had not been explicitly trained. Sidman proposed that the three stimuli —printed word, spoken word, and drawing— were now members of a category and were now "equivalent" to one another.

The emergence of a relation that was not explicitly trained was interesting because it did not yield to an explanation using existing behavioural principles. Behavioural researchers were already familiar with the concept of a conditional discrimination, in which an organism is taught to emit a certain response in the presence of a certain stimulus (in a sort of "if-then" relation), but neither conditional discrimination nor any other known behavioural principle could account for the spontaneous formation of the "printed-drawing" relation that had not been trained.

Once stimuli become members of a single category of stimuli, behavioural performances in tasks involving these members show interesting properties. In abstract terms, if a human participant is trained to match A to B and A to C in a series of conditional discrimination tasks, that individual may match B to A and C to A, and B to C and C to B, without any specific training to do so. When this pattern of spontaneous matching responses

emerges, Sidman suggested that we define them as participating in, or forming, an equivalence class or relation.

Inspired by set theory in mathematics, Sidman argued that equivalence relations had three defining properties. The first property was reflexivity and was demonstrated when a participant matched each stimulus to itself (A-A, B-B, and C-C). The second property was symmetry and was shown when participants spontaneously reversed each trained matching response (A-B yielded B-A matching, and A-C yielded C-A matching). The third property was transitivity and was shown when a participant spontaneously combined the trained relations across a mediating node (A-B and A-C matching yielded B-C and C-B matching). Note, that in the latter case, the performance would actually be defined as combined symmetry and transitivity because it apparently involved both properties (see Sidman, 1994 for a detailed description).

The phenomenon of stimulus equivalence was explored extensively over the next few years in numerous studies that found that it appeared readily in the behaviour of verbally-able humans but not so readily in non-humans (if at all) or severely language impaired humans (Devany, Hayes, & Nelson, 1986). And despite on-going efforts over the next 30 years or so there is still very limited evidence for the most basic forms of equivalence class formation in non-human participants (Dymond, 2014; Hughes & Barnes-Holmes, 2014)

Excitement over the equivalence phenomenon built up quickly due to its apparent overlap with symbolic relations in natural language (for an interesting exchange on this area see Sidman, 1994; letters between Murray Sidman and Willard Day). In general terms, the link between stimulus equivalence and human language, or at least symbolic relations, was widely recognised within the behavioural research community. However, three different conceptual perspectives on the nature of this relationship emerged during the late 1980's and early 1990s. In brief, Sidman suggested that stimulus equivalence should be considered a

basic behavioural process that may account for the symbolic properties of human language. In contrast, other behavioural researchers suggested that human language, and in particular, naming served as the basis for the formation of equivalence relations (Dugdale & Lowe, 1990). The third conceptual approach that emerged during this period extended beyond an account of stimulus equivalence and/or symbolic relations *per se,* and instead used the work as a "spring-board" to develop a broad and ambitious theory of human language and cognition. This latter perspective quickly gained momentum as a research programme during the late 1980s and throughout the 1990s leading in 2001 to a full book-length treatment, entitled *Relational Frame Theory: A Post-Skinnerian Account of Human Language and Cognition.*

Relational Frame Theory

In contrast to Skinner's (1957) treatment of human language, RFT was developed specifically to generate a systematic programme of research in this domain. As such, RFT is a contemporary behaviour analytic approach to human language and cognition, which aims to offer a comprehensive, empirically-supported, functional-analytic account, including a treatment of moral behaviour (Hayes et al., 2001).

RFT integrates a number of key conceptual and empirical developments within behaviour analysis. The first of those pillars is the ability, found in many different species, to respond relationally based on the formal properties of the relevant stimuli. Selecting the smaller, or darker, or wider object in a simple discrimination task provides a simple example. In effect, with appropriate training, many complex organisms are able to respond in the presence of stimuli they have not previously encountered, but whose physical properties (dimensions, colour, brightness, etc.) may be used to control a particular pattern of relational responding. Although this is true for most species, relational responding in humans is not limited to physical properties. For instance, a person can be presented with a number of

different objects and asked to select the most valuable, but value is a non-physical construct, an abstract property that may be based on social whim (e.g., a 5 euro note is more valuable than a 10 euro note).

According to RFT, the ability to respond relationally based on contextual cues that extend beyond the physical properties of the to-be-related stimuli is explained by appealing to the concept of a "generalised", "purely-functional" or "overarching" operant response class (see below). The basic idea is that exposure to the verbal contingencies operating in the natural environment of most humans provides literally thousands of exemplars of reinforced relational responses in the context of specific contextual cues. For example, a young child may be exposed literally to hundreds of thousands of "naming" exemplars during the first few years of life, in which specific cues come to predict the bi-directional nature of symbolic relations. Questions, such as "Is this your mommy?" and "Is this your teddy?" and so on, serve to establish the word "is" and the naming context more generally, as one that predicts reinforcement for a two-way relationship between symbol and object.

In this case, if the child hears the word teddy and then orients towards the actual toy, social reinforcement may follow (e.g., smiling and praise from the care-giver). Furthermore, if the child looks at her teddy on another occasion and then smiles and giggles if the caregiver says "Are you looking at your teddy?" then again social reinforcement may follow. RFT suggests that exposure to many such examples serves to establish a relational operant that is controlled by specific contextual cues (in this case the word "is"). With sufficient exposure across a large enough number of exemplars, "training" in one direction (look at object-hear word) may generate the spontaneous emergence of relational responding in the opposite direction (hear word-look at object), without having to provide direct reinforcement, instruction or prompting beyond the presence of the relevant contextual cue.

The same general logic is applied to explain the emergence of a wide range of patterns of relational responding, which are labelled *relational frames*. For example, the words "bigger" and "smaller" may come to function as contextual cues following exposure to a sufficient number of relevant exemplars. For illustrative purposes, imagine a young child who is told that a "dog is smaller than a horse". If she were asked subsequently, "Is a horse smaller than a dog," she may answer yes, and in this case she has failed to derive what RFT refers to as the frame of comparison between the two words (horse and dog). In the natural language environment, of course, it is likely that a care-giver would correct the child's response in this instance and say, "No, a horse is bigger than a dog, not smaller". Gradually, across many such exemplars the contextual cues ("bigger" and "smaller") will come to control appropriate relational responding. Thus, if a child is told that X is bigger than Y, they will spontaneously derive that Y is smaller than X without further instruction, reinforcement or prompting. These types of learning histories are referred to as generalized or over-arching operant classes because the history involves generalizing across (or arching over) many exemplars before the final operant pattern of relational framing itself is established in the behaviour of the young child.

Contextual events and conditions can be functionally tied to specific types of relational frames, so they are able to initiate particular relational responding patterns. Just like equivalence relations, relational frames have certain properties, some of which resemble those of equivalence. These are mutual entailment (roughly symmetry), combinatorial entailment (roughly transitivity) and the transformation of stimulus functions. The third property refers to the acquisition of psychological functions by virtue of participation in a relational frame. The transformation of stimulus functions thus helps to explain how symbols and other verbal stimuli, such as abstract concepts, come to elicit emotional responses. The concept of the transformation of functions is central to RFT and thus it will be described in detail here.

Suppose that a young boy was attacked by a dog and as subsequently experiences a high level of fear whenever he is approached by a dog. Simply telling the child something like "we are going to visit a relative who has a pet dog" may also evoke a similar state of anxiety and fear, although no actual dog has been observed in the present moment. The effect appears to involve more than classical or Pavlovian conditioning, in that relational framing is involved. That is, having learned to frame events relationally, as a generalized operant pattern of behaviour, if actual dogs enter into a frame of coordination with the word "dog", the latter may now be *transformed* into an aversive or "fear-inducing" stimulus in and of itself.

Critically, participation in relational frames allow for the emergence of different patterns of transformations of functions depending on the type of frame. For instance, the positive functions linked to a certain stimulus may become negative if the said stimulus participates in a relational frame of distinction. A relevant example would be a situation in which a person is presented with a previously unknown animal, in this case a ferret, and is told that it is "not at all aggressive or dangerous." Insofar as the phrase "not at all" functions as a contextual cue for distinction, the ferret may acquire approach rather than avoidance functions for the listener. Or to put it more informally, the ferret is seen as relatively safe to approach because "safe" is in a frame of distinction with "aggressive" and "dangerous".

The transformation of stimulus functions may also be involved in instances of behaviour in which partial information serves to evoke psychological functions. A mother who says "An engineer? I like her already" when her son talks about his new girlfriend, shows that psychological functions have been already linked to an uncontacted stimulus by virtue of its participation in a relational network (in this case the label "engineer" with well "educated" and "professional"). This process, almost instantaneous in many different instances of human behaviour, is possible because of relational entailments and transformation of

stimulus functions (i.e., relational framing). Through interactions and experience (continued relational framing), individuals are able to build increasingly complex relational networks, full of overlapping concepts, actions, and psychological functions. Eventually, the verbal repertoire becomes so sophisticated that an individual can operate (i.e., frame relationally) in abstract worlds that have few, if any, physical properties directly accessible to the organism, such as mathematics, logic, perspective-taking, and future plans. The types of relational framing involved in the construction of such complex relational histories will be described now.

Families of relational frames and complex relational networks.

Several different types or families of relational frames have been identified. Developmentally, the earliest ones are probably those of coordination and distinction. Coordinative relational framing conveys the sense of sameness or general equivalence, and it is evoked by contextual cues such as the utterance of the word "is"; for instance, when a parent points at a dog and tells a child at the same time "that is a dog". The contextual cue prompts the relational response of coordination between the animal and the word "dog". Responding in a frame of distinction, on the other hand, is probably controlled by utterances such as "is not" or simply "is different from". In addition to coordination and distinction other relational frames also appear in the behavioural repertoires of young children, such as opposition, comparison and class containment or hierarchical frames. Finally, deictic relational frames have been widely discussed and studied. These frames involve relating a speaker to others (as in I versus You) and locating speakers and others in time and space. For example, young children learn through interactions with the verbal community to utter and understand statements such as I am here (at home) now (at the current time), but I was there (at school) then (an hour ago).

Relational frames typically involve only involve only three stimuli or events. One example would be a coordination relation between the word "dog", the onomatopoeic "woof" and the picture of a dog, which naturally involves mutual and combinatorial entailments and transformation of functions. However, families of relational frames can combine to form complex relational networks. For example, a person can give another the following complex instruction: "I will leave on holidays in two weeks and will be gone for a month. If you water and mow my lawn each week I am gone, the following month I will pay you $100." (Hayes et al., 1998, p. 256) The ability to respond in terms of several core relational frame families is required to understand and follow the instruction: before-after frames ("mow *after* two weeks"), if-then frames ("*if* you do it *then* you will get money"), and even basic coordination frames ("grass" with certain classes of physical events).

Complex relational networks involving deictic frames, it has been argued, are the building blocks of a sense of self. They allow people to respond accurately to questions such as "What did *I* do?" or "Where were *you then*?". When combining with conditional frames, relational performances involving moral components may start to appear: a question such as "What *would you* do if *you were him*?" is complex in that it requires the ability to respond in accordance with several relational frames in order to be understood. As the developing person has more opportunities to frame relationally in terms of perspective, a sense of self consolidates and acquires more and more properties. In RFT terms, this amounts to increased complexity of relational networks involving perspective.

Observing Relational Framing: the Implicit Relational Assessment Procedure

Supporting evidence was available for the general principles of RFT upon its formal presentation in 2001, but there was no way of observing relational framing "on-the-fly" until the mid-2000's, when the Implicit Relational Assessment Procedure (IRAP) was designed (Barnes-Holmes et al., 2006; Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008). The

IRAP is a computer-based task designed to present a context in which the strength and directionality of relational framing can be assessed (Hussey, Barnes-Holmes, & Barnes-Holmes, 2015).

The IRAP is based on the idea that across time, well established relational responses tend to be emitted faster and more accurately in evocative contexts, whereas less well established responses may be emitted at relatively lower levels of accuracy and speed. The former (well established) responses have been described as *Brief and Immediate Relational Responses* (BIRRs for short) and the latter as *Extended and Elaborated Relational Responses* (EERRs). Even though it may be tempting to assign the more traditional labels of "automatic" vs. "controlled" processes to both, the use of the behavioural terminology is a reminder of a different, functional conceptual framework upon which an RFT account of such responding rests. The formal model is called the Relational Elaboration and Coherence (REC) model (Barnes-Holmes, Barnes-Holmes, et al., 2010; Hughes et al., 2012), and its basic premise is as follows.

When an individual first acquires a particular relational response it may be more EERR like (than BIRR like), particularly if it involves deriving a particular relation. Thus, for example, if a person learns that A is the same as B and B is the same as C, initial responses that involve relating A and C as the same may be considered relatively high in derivation and complexity. That is, the person may work through a relatively complex derived response, such as "if A is the same as B and B is the same as C, then A and C must be the same". If, however, the person is presented with many opportunities to derive this relational frame, the level of derivation and complexity will likely decline, such that the person may come to emit the simple relational response "A same as C." At this point, the relational response is more properly considered a BIRR. The IRAP was specifically designed to capture the more BIRR-like properties of relational responses than are likely to be reflected in other measures of

relational responding. This is achieved primarily by asking participants to emit specific patterns of relatively simple relational responses under time pressure (i.e. requiring brief and immediate responses).

A typical IRAP presents six test blocks preceded by a variable number of practice block pairs, each block consisting of 24 trials. Each IRAP trial is presented on a computer screen and requires participants to indicate, quickly and accurately, the relationship between a label stimulus presented at the top of the screen, and a target stimulus presented below the label. Response options that may indicate the relationship (e.g., "same" and "opposite") between the label and target stimuli are presented at the bottom left and bottom right of the screen. In some IRAPs their left-right positions are randomised across trials. Each block of trials requires that participants respond in accordance with one of two patterns that are deemed to be consistent or inconsistent with a particular response bias. Thus, one block of trials might require that participants respond in a manner that reflects natural verbal categories and another block would require the orthogonal pattern. For example, if the label and target stimuli, "Pleasant" and "Love" were presented in a consistent block, choosing the response option "Similar" would be deemed the correct response, but during an inconsistent block of trials choosing "Opposite" would be deemed correct. The typical IRAP consists of four different trial-types, based on pairing each label with each target stimulus, with each trial-type being presented an equal number of times within each block (see Figure 1). The IRAP program presents the two types of blocks of trials (consistent versus inconsistent) in an alternating pattern throughout the practice and test phases of the procedure.

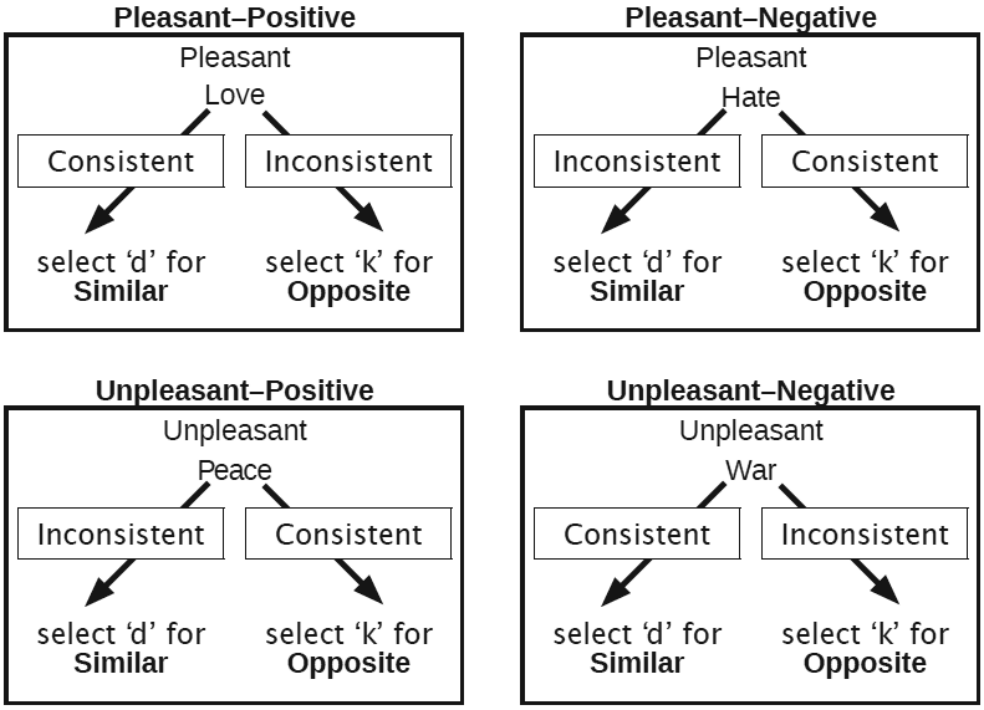


Figure 1. An example of the four trial types in the IRAP (source: Barnes-Holmes, Barnes-Holmes, et al., 2010, p. 531)

If a response is not made on any given trial within the specified time criterion, the program displays a customisable "Too Slow" warning in red in order to advise the participant that faster responding is needed. When the response is incorrect (consistent response in inconsistent blocks or inconsistent response in consistent blocks), a red X is displayed and the participant must enter the correct response in order to proceed to the next trial. A guided introduction with practice blocks is performed by the researcher before the actual test blocks, which are only presented if the participant achieves predefined accuracy and latency criteria in the final pair of practice blocks (normally >=80% correct responses and <=2000 milliseconds average latency).

The presentation software outputs response accuracy and latencies for each trial, which are analysed according to a standardised protocol which transforms accuracy/latency data into D-scores using a version of the D algorithm by Greenwald, Nosek and Banaji (2003),

called the D-IRAP score, which seems to control for individual differences such as cognitive ability, age and other factors (Barnes-Holmes, Murphy, Barnes-Holmes, Stewart, & Boles, 2010).

The IRAP has been widely used in several assessment contexts with promising results. It has proven useful in assessing implicit cognitions in a range of domains, from fear (Nicholson & Barnes-Holmes, 2012b) to eating disorders and perceptions of body image (Parling, Cernvall, Stewart, Barnes-Holmes, & Ghaderi, 2012; Roddy et al., 2011), to the prediction of drug treatment outcomes (Carpenter, Martinez, Vadhan, Barnes-Holmes, & Nunes, 2012). It seems especially suited to the assessment of complex sociocultural issues such as prejudice and morals because of its functional roots and its ability to tap into the directionality of the relationships between concepts. Indeed, a recent meta-analysis indicates that the IRAP predicts clinically relevant criterion variables at $r = .45$, which compares favourably with all other measures of implicit cognition, including the IAT (Vahey, Nicholson, & Barnes-Holmes, 2015).

The research reported in the current thesis draws heavily upon RFT, the REC model and the IRAP in an effort to develop a behaviour-analytic approach to the study of moral behaviour and, so called, cheating responses in particular. Before concluding the current chapter, we will consider how RFT approaches the topic of morality and cheating behaviours, and consider some of the experimental procedures that have been developed to examine cheating itself under laboratory conditions.

A Relational Frame Account of Morality

By explaining how humans can dispense with the need for direct training of a large number of elements in their learning histories by deriving relations and transforming psychological functions, RFT provides a conceptual framework for studying human

behaviour that would typically be seen as difficult to explain in terms of direct histories of reinforcement or generalization performances that are closely tied to the physical properties of the environment. As we shall see, some features or properties of moral behaviour appear difficult to explain in terms direct acting contingencies and physical generalization processes, and this is why RFT seems to be well-suited to study morality.

Relational Frame Theory is firmly grounded in behaviour analysis, and thus a view of moral behaviour from this perspective retains foundational premises of the operant account, such as the cultural/social dependence of human morals and rights, the rejection of internal causative agents, and the recognition of the need for a scientific study of value judgments. It adds to this account the role of language and an explanation of the formation of increasingly complex and overlapping relational networks, that can be used to functionally explain the coherence, and lack thereof, in human moral behaviour. In RFT, moral behaviour is hence defined functionally as "behaviour governed by, and consistent with, verbal rules about what is socially and personally good" (Hayes & Hayes, 1994, p. 46). Therefore, a small review of rules and what they mean in behaviour analysis and RFT specifically is in order before moving on.

Rules and rule-following in RFT

The notion of rules has been used before in behavioural psychology to explain the kind of complex behaviour in which humans respond in consistent ways in the absence of a direct training history, in what has been called rule-governed behaviour, as opposed to contingency-shaped behaviour. The concept of rule-following was first advanced by Skinner in his approach to problem solving (1966), where rule-governed behaviour was defined as behaviour under the control of stimuli that specify contingencies, known as instructions or rules.

Although there was some value in this proposal, it failed to provide an adequate functional explanation of rule-following itself. For example, Skinner argued that rules specify contingencies but he failed to articulate exactly what it means to specify a contingency in functional terms (for a detailed discussion see Hayes et al., 2001). Only with the development of RFT did such an account become possible, since it conceptualises rules as verbal stimuli that participate in relational frames (Hayes et al., 2001). As noted earlier, different classes or patterns of relational framing may combine to form complex relational networks that may function as instructions or rules (e.g., "I will leave on holidays in two weeks and will be gone for a month. If you water and mow my lawn each week I am gone, the following month I will pay you $100."; Hayes et al., 1998, p. 256). The ability to understand this instruction involves before-after frames ("mow *after* two weeks"), if-then frames ("*if* you do it *then* you will get money"), and basic coordination frames ("grass" with certain classes of physical events), and the appropriate transformations of functions in accordance with those frames. For example, the statement "it is three weeks since I was asked to mow the lawn" may now function as a verbal stimulus that evokes lawn-mowing in the listener (because the speaker has been on holiday for two weeks and a week has passed since she left). On balance, as most parents of teenage children know, understanding a rule does not automatically mean that it will be followed. Nevertheless, the need to provide a functional analysis of rule-following, rather than just rule-understanding, has also been addressed within RFT.

Hayes et al. (1998) proposed that two functional classes of rule-following behaviour can be established depending on the nature of the controlling history. In the case of *pliance*, rule-following behaviour is under the control of socially mediated consequences for the correspondence between the rule and the relevant behaviour. For example, reducing speed upon seeing a sign that says "Maximum speed: 50 km/h" is probably under the control of positive consequences for following that type of rule (traffic signs) and negative consequences for their transgressions. In this case, failing to follow the rule may lead to

punishment by a law-enforcement office or agency. In other words, the listener complies (hence “pliance”) with the rule because doing so is consequated by the “rule-giver.”

A second functional class of rule-following is *tracking*, where the source of control is a correspondence between the rule and the contingency it describes. For instance, a rule such as "the plate is hot, please be careful” will likely be followed because of a history of reinforcement or punishment for following or failing to follow such rules in the past. In this case, the consequences of rule-following are provided by the physical environment rather than by the rule-giver (i.e., following the rule ensures that the listener is not burnt by the hot plate). In effect, the rule is a *track* or guidepost that specifies a contingency that is independent from the consequating behaviours of the individual who provided the rule.

Pliance and tracking can be affected by a process called *augmenting*, which involves the creation of new consequences (formative augmenting) or the enhancement of pre-existing ones (motivative augmenting) through the use of *augmentals*, verbal stimuli that change the relative strength of reinforcers or punishers in a contingent relation. An example of a formative augmental could be “This voucher may be exchanged for a free item in the supermarket”. That is, the statement may establish reinforcing functions for the voucher, without any direct history of reinforcement with the voucher or physically similar stimuli. Motivative augmentals are seen as altering the reinforcing or punishing functions of particular stimuli. For example, the statement, “It’s hot -- wouldn’t it be good to drink a glass of *cold, foamy, refreshing* beer?” may serve to increase the reinforcing functions of beer, such that a listener responds in a manner that allows access to the reinforcer (e.g., by driving to a supermarket to buy some beer). Critically, the statement does not increase or decrease the availability of beer (the listener was free to drive to the supermarket at any time). Rather, simply hearing the rhetorical question causes the listener to seek out beer at a higher

probability than if he or she had not heard the question. It appears that a great deal of advertising, at least for "cash-cow" products, is based on this psychological process.

Hayes et al. (1998) suggest that children learn pliance first because it is convenient for the verbal community and because it lays the foundation for higher sophistication in rule-following. Tracking then helps the child make effective contact with the way consequences are arranged in the natural environment. Augmenting makes it possible to increase or decrease responding that facilitates access to novel consequences (formative augmentals) or to manipulate the extent to which previously experienced consequences are established as reinforcing or punishing at a particular point in time (e.g., an ice-cream on a hot summer's day).

Relational framing in moral issues

To illustrate how verbal or relational responding appears to be central to moral behaviour, consider this example by Haidt and Joseph (2004): two biological siblings agree to have protected sexual intercourse, and despite not regretting it and finding that it strengthened, instead of undermining, their personal relationship, they decide to keep it secret and never to do it again. Most people will condemn this behaviour as immoral, even though the usual reasons of the genetic dangers of incest or psychological trauma are not present, ultimately stating that they are not sure of the reason, but they simply know it is morally wrong. However, if we were to simply substitute the word "siblings" with "strangers", the moral acceptability of this situation would change, at least for some people. In this case, participants would be responding "wrong" when they are asked about the *relation* between "sex" and "siblings", but "right" when the relation is between "sex" and "strangers".

In effect, the reaction that the majority of readers will have to the example of siblings having sex will be strongly negative, not because they will have had any direct experience of incestuous relations, and the potential negative consequences, but because "incest" participates in a rich network of verbal relations that serve to establish relatively strong negative (or taboo) functions for the act itself. Moral judgements, therefore, do not require direct experience or exposure to relevant contingencies of reinforcement and/or punishment, but the establishment of increasingly complex and rich relational networks that serve to establish specific actions as either "moral" or "immoral". Ultimately, moral judgements that are largely verbal may be traced to perhaps directly experienced events (increased chances of genetic abnormalities in the off-spring produced by siblings), but such events were likely experienced by our ancestors in the distant past, and only very rarely by humans in modern culture today. In other words, our moral aversion to incest may be largely verbal, not experiential.

Classical moral dilemmas offer an opportunity to see how these verbal processes underlie what has been traditionally called "moral reasoning". Consider the trolley problem mentioned earlier. Specifically, participants frequently confirm that they would "flick a switch" that would lead to the death of one person but save the lives of five others; however, far fewer individuals confirm that they would physically push a person under the trolley if doing so saved five other lives. From an RFT perspective, the cost-benefit ratio of the network is identical across the two examples (losing one life saves five), but the psychological functions evoked by the two scenarios are dramatically different. That is, flicking a switch (remotely) is less likely to elicit the highly negatively valenced functions of physically pushing another human being under a train carriage.

The foregoing example begs the question as to why sacrificing one person in two different ways (flicking a switch versus physically pushing) evokes such contrasting

transformations of functions. One answer to this question would focus on the importance of deictic framing in moral reasoning and decision making. As noted earlier, deictic relations are involved in learning to engage in perspective-taking. Through this type of learning, a sense of self versus others emerges. In the early years, the perspective-taking may be relatively simple, as when a child learns to report what he or she is eating versus what someone else is eating during a meal. As this relatively basic type of deictic relational responding becomes extended and elaborated across many different contexts, specific moral codes or rules may be specified by care-givers. For example, the advice, "Do unto others as you would have done unto you" requires that the listener engage in relatively complex relational responding to first establish how he or she would like to be treated by others and then to apply that to his or her treatment of other people. Technically, this requires a type reversal in I-YOU relations of the form, "if I was you and you were me, how would I feel if you did X to me". If the answer is "I would feel bad" then the moral code requires that you do not do X to other people.

Note, however, that this is not simply an abstract relational issue. Although reversing the I-YOU relation is required, the verbal action also requires that the negatively valenced functions of X are evoked for the individual and these are then transferred to the other person. The maxim thus requires what might be described as a type of verbal "empathy". In other words, deictic framing helps us as individuals to feel the pain of others based on transformations of psychological functions among complex relational networks that include the ability to engage in the reversal of deictic relations. Critically, this ability ensures that verbal morality is not simply a matter of following abstract (purely relational) rules, but involves, in a verbal sense, experiencing the pain and suffering of others.

Thus, when a person engages in an immoral act that would cause pain and suffering to another human being, it may well elicit or evoke some level of pain or suffering in the perpetrator too. Insofar as this is the case, flicking a switch from a remote location that

causes the death of another human being (in which the actual death is not witnessed) may well be seen as far less aversive than physically pushing a person under the trolley and being forced to witness the actual death of another human being. Metaphorically, in killing another human being, verbally, one may kill oneself (or less metaphorically induce a sense of guilt so strong that one is unable to live with it).

Summary

The specification of behavioural guidelines or rules that serve certain social purposes is central to morality. From the point of view of RFT, rules can be conceptualised as relational networks that people come to understand through continued opportunities to exercise different types of relational framing, and then follow through pliance, tracking and augmenting. But morality also incorporates an emotional component that can be accounted for from the point of view of RFT through transformations of functions via deictic relations. Those two pillars, transformation of functions and rule understanding and following, enable RFT to provide an account of morality that reconciles the cognitive and emotional dimensions of moral behaviour in a unified framework.

This conceptual view of morality has been part of RFT since the beginning (Hayes et al., 2001), but less well established from this theoretical perspective is that responses relevant to human morality may occur relatively slowly or rapidly and the resulting response classes may be functionally distinct. For example, moral relational responses that are slow and deliberate may come under the control of extraneous social variables such as social desirability, whereas fast responses may be less susceptible in this regard. Indeed, as noted earlier the basic argument of this theoretical position, has been articulated formally in the context of the REC model (Barnes-Holmes, Barnes-Holmes, et al., 2010).

It is worth noting that the concept of relatively fast relational or verbal responding, highlighted by the REC model, shows potential overlap with other theoretical approaches. For example, the Social Intuitionist Model of Ethics (Haidt, 2001), is a cognitive model that claims that many if not most moral evaluations or judgments made with respect to culturally-dependent virtues come from quick moral intuitions that are then followed, if needed, by moral reasoning. The parallels with the behavioural account that we have described are easy to see, in the form of socially-mediated verbal histories, which involve brief versus extended relational responding under relevant forms of contextual control.

At the time writing the author was unaware of any published (or unpublished) research that had attempted to study moral behaviour specifically from an RFT perspective that also drew on the recent developments with the IRAP and the REC model. Nicholson and Barnes-Holmes (2012a) used a socio-moral task within the context of developing an IRAP to measure disgust, in which participants were asked to think about moral violations and to rate feelings evoked by the thought of transgressing them, but the study did not intend to focus specifically on the IRAP as a predictor of moral choice or feelings related to moral decisions.

The overarching or general aim of the research reported in the current thesis was to lay the groundwork for this empirical investigation. In pursuing this line of inquiry there were many possible ways of attempting to capture behaviours in an experimental context that could be seen as involving an important moral dimension. Indeed, there is a reasonably well developed literature on various laboratory-based tasks and procedures that have been used to assess or measure moral versus immoral behaviours in the form of "cheating" tasks. Specifically, these tasks typically present research participants with an opportunity to engage in an act or acts that involve deception or lying in some way that gains some advantage for the perpetrator. In effect, participants are placed in a context in which they may behave

morally (choosing not to cheat on a task) or immorally (choosing to cheat). The current research drew heavily on this work.

Chapter 3

Introduction to the Current Research Programme

Introduction to the Current Research Programme

As noted at the end of the previous chapter, a range of different tasks have been employed in the psychological literature to study immoral behaviour in experimental settings, operationalised as "cheating". Most of them involve discreetly providing an opportunity to do better on a task, especially if the result is paired with monetary compensation. For example, giving a discreet chance to consult a dictionary during a vocabulary test, where correct responses were contingent on monetary payoffs(Ong & Weiss, 2000, p. 1695-1699), or receiving a small amount of money for finding pairs of numbers in a timed visual search task (Shu et al., 2011; Vohs & Schooler, 2008), where it was possible for participants to over-report the actual number of identified number pairs.

In the current programme of research, it was decided to use the Mental Math Task (Von Hippel, Lakin, & Shakarchi, 2005), except for one experiment (the rationale for which will be explained later in the thesis). In this computer-based task, participants have to complete two sets of 10 equations consisting of numbers from 1 to 20 to be subtracted and added. The opportunity to cheat is provided by a putative bug in the program, which allows the participants to see the correct answer to each problem. Unbeknownst to the participants, the program logs the number of trials in which the program bug is used to cheat by each participant.

Several reasons prompted us to use this task for most of the experiments: (i) it is the most widely used task in the literature to operationalise cheating behaviour; (ii) it tends to produce relatively high amounts of cheating; (iii) it can be employed within a single session; (iv) it is relatively simple and straightforward to perform; (v) in contrast to some other measures of cheating it provides an individual score for number of cheats for each participant, and (vi) it is largely independent of verbal ability, since it only involves basic arithmetic. Furthermore, pilot work indicated that it was relatively easy to employ and

worked reasonably well with the type of sample population (college students) that would be employed in the current research.

A brief description of the entire research programme will follow, before presenting the individual studies in chapters 4-6.

Description of the Studies

To begin this research programme, we intended to determine whether specially-tailored IRAPs correlate with performance on a deception task, and could therefore be used as predictors of moral behaviour. In Chapter 4, we report two studies where we use two IRAPs with different measures of deception and moral disengagement, in an initial exploration of relational framing involved in immoral behaviour. The two IRAPs were designed to tap into beliefs about what is good and bad, and also about the amount of moral or immoral behaviour emitted by participants on a daily basis. The key difference between these two exploratory studies was the use of different measures of deception.

In this first experimental phase, we confirmed conclusions from previous research, namely that participants readily classify good and bad actions, and that they rapidly state that their behaviour is often good. However, some of the results support the multi-dimensionality of morality and the existence of grey areas where moral opinion seems to depart from commonly held values and virtues.

Encouraged by the results from this first phase, we decided to explore the role of psychopathy, which seems to predict moral choice (Tassy, Deruelle, Mancini, Leistedt, & Wicker, 2013) on moral behaviour in student samples from Ireland and Colombia, to see also if the IRAP reflects any cultural differences. The two studies are presented in Chapter 5. We designed and tested another IRAP in this phase, intended to tap into positive or negative

feelings upon engaging in moral and immoral actions. The main finding was a stable correlation between certain trial types in this latter IRAP and the cheating measure.

The final experimental phase consisted of two separate studies and focused on a values-oriented intervention on immoral behaviour. The intervention stems from Terror Management Theory (Solomon, Greenberg, & Pyszczynski, 1991) which proposes that cultures as symbolic systems help people give meaning to their lives and alleviate the distress created by the inevitability and proximity of death. The procedure itself, called the Mortality Salience Intervention (MSI), features two different components, one that focuses on the relative shortness of life (the time component) and another that focuses on the feelings evoked by the inevitability of death (the mortality component proper). In study 5, we found a significant effect of the MSI on the moral feelings IRAP and on cheating, and in study 6 we separated the components of the MSI to examine the contribution of each to the effect observed.

In the seventh and final chapter, a summary of the research is provided and a range of empirical and conceptual issues stemming from the empirical studies are discussed.

Ethical Considerations for this Research Programme

As mentioned before, both anecdotal and empirical evidence coming from different fields of psychology has identified that people are prone to distort their answers in interviews and questionnaires in order to present themselves as possessing desirable attributes. This is generally known within psychology as "Social Desirability", and implies that people value certain behavioural dispositions as good and others as bad. It has also been observed that people seem to think that good or bad behaviour is the reflection of internal, relatively stable inclinations, so they readily label persons, rather than behaviours, as "good" or "evil" – this is called the "fundamental attribution error" (Gilbert & Malone, 1995).

Naturally, facing participants with results that suggested that their behaviour does not quite match their impressions of it could potentially create a state of mild psychological distress with anxiety and uncertainty. We identified this and other potential concerns raised by the nature of our research and aimed to address them according to the general principles laid out in the Code of Ethics and Conduct of the British Psychological Society (2009), the Guidelines for Safe Work Practice of the Department of Psychology (2015) and the Ethics and Deontology Code of the Colombian College of Psychologists (2006). Specifically, we paid careful attention to the following considerations:

a) A small monetary compensation of €5.00 (or roughly equivalent $10.000 COP) was provided to all participants throughout the studies. Participants were given this financial compensation right after signing the informed consent, and at the same time they were told that it was theirs to keep from that moment without prejudice to their rights as participants, specifically their right to withdraw from the experiment at any time.
b) Full debriefing was performed at the end of every experimental session, specifically involving a thorough explanation of the true purpose of the deception measures and the rationale behind the justified deception (reactivity due to social desirability). Critically, participants were also told that there was nothing inherently good or bad about their answers, and that the tasks were not personality tests. Concerns were addressed carefully to ensure that the participants left the experimental session in a positive psychological state.
c) A protocol was in place to deal with manifestations of heightened distress as per the Guidelines for Safe Work Practice of the Department of Psychology (National University of Ireland - Maynooth - Department of Psychology, 2015), involving the termination of experimental tasks and immediate remission to the University Medical Centre. Fortunately, at no point did this protocol need to be activated.

Chapter 4

An Initial Exploration of Morality with the IRAP

Study 1: Investigating the Use of the IRAP as a Predictor of Cheating

In the first study we started our exploration of the IRAP as an estimator of the probability of engaging in immoral behaviour. With this in mind, we used a well-known cheating task (Von Hippel et al., 2005) as an operationalisation of immoral behaviour, and set out to determine if a set of IRAPs could measure the likelihood of engaging in cheating.

We hypothesised that two different types of relational responding could be involved in this sort of moral decision making. The first is the ability to classify actions as good or bad, independently of the moral actor who performs the action. In order to capture this type of response we designed an IRAP that we called the "Conceptual Morality" (CM) IRAP. This IRAP asks participants whether good or bad actions are in fact good or bad, and serves as a starting point to begin exploring relational networks related to morality.

As discussed previously, RFT suggests that a key component of morality involves deictic relational responding. Specifically, a given individual may recognise that certain actions are perceived to be immoral by the wider culture but not necessarily agree with those views. With the goal of tapping into the more deictic properties of moral responding, we also employed what we called a "Deictic Morality" (DM) IRAP. The primary goal of this first study in the current research programme was to determine if performance on one or both of the two IRAPs predicted performance on the measure of cheating.

Previous studies have reported that people tend to provide explanations for their wrongdoings as a way of reconciling their positive self-image with the negative assessment of the morality of their actions (Bandura, 2002). This process is called *moral disengagement* in Social Cognitive Theory, we hypothesised that it may be conceptualised as a verbal process and accounted for from an RFT perspective. Therefore, we included measures of moral

disengagement to determine if the IRAP(s) could capture relational framing related to a tendency to justify immorality.

Material and Methods

Participants. A convenience sample of 38 students (64.9% females) with ages between 18 and 35 (M = 22.43, SD = 4.28) from the National University of Ireland, Maynooth, volunteered for this study. One participant decided to withdraw from the study as he was completing the IRAP, so he was thanked, debriefed and dismissed, and another nine did not reach the test criteria in at least one of the IRAPs – the data for these ten participants were excluded from analysis, leaving a sample consisting of 28 participants. All participants had a high level of fluency in English.

Measures. Cheating was operationalised as the performance on the *Mental Math Task* (Von Hippel et al., 2005). This is a computer based task that presents participants with two sets of 10 equations, each consisting of ten numbers between 1 and 20 to be mentally (i.e., without using a calculator or pen and paper) subtracted and added. Each equation is preceded by a prompt that reads "Here comes the next one" for 500 milliseconds. In order to enter an answer, the user must press the Spacebar within the allotted time of 10 seconds for the first set ("Slow block") and one second for the second set ("Fast block") to make an input box appear. If the participant does not press the Spacebar within that period, the correct answer shows up below the equation, but the input box can still be brought up to enter an answer. The program logs whether the participant waited to see the answer, which constitutes cheating. This task has been used in other studies with slight modifications (R. P. Brown et al., 2011; R. P. Brown, Budzek, & Tamborski, 2009; Jordan, Mullen, & Murnighan, 2011; Shariff & Norenzayan, 2011; Teper & Inzlicht, 2010; Vohs & Schooler, 2008). The task is preceded by instructions designed to give the impression that the possibility of seeing the correct answer is a "bug" in the program, and participants are specifically asked to press the

Spacebar as soon as they see the equation appear, which will prevent the correct answer from appearing. The original software that we used was kindly provided by its author.

The *Civic Moral Disengagement Scale* (Caprara, Fida, Vecchione, Tramontano, & Barbaranelli, 2009) was designed to assess moral disengagement in the realm of civic duties, that is, behaviours related to the use of public resources and property. The 32-item version used here has a Cronbach's alpha of 0.92. The questionnaire is scored in a range of 32 to 160 using 5-point Likert scales, with higher values indicating greater disengagement (see Appendix A). Items are grouped as part of eight conceptualised mechanisms of moral disengagement: Moral justification, Euphemistic language, Advantageous Comparison, Displacement of Responsibility, Diffusion of Responsibility, Distorting Consequences, Attribution of Blame and Dehumanisation.

The *Moral Disengagement Scale* (Bandura et al., 1996; Bandura, 2002) is a 5-point, 33-item questionnaire designed to assess levels of moral disengagement. Despite being originally conceived for school-aged children, it has been used with undergraduate (e.g. Jackson & Gaertner, 2010) and adult (e.g. Claybourn, 2010) populations. Participants in the pilot study had difficulty understanding item 15 ("It is okay to treat badly somebody who behaved like a 'worm'"), so we changed it to "If someone acts like a jerk, it is okay to treat them badly", as suggested by Pelton, Gound, Forehand and Brody (2004). The authors reported a Cronbach's alpha of 0.82 and the complete scale is reproduced in Appendix B.

In order to assess social desirability bias, the widely-used *Marlowe-Crowne Social Desirability Scale* (Crowne & Marlowe, 1960) was employed. The scale is a 33-item questionnaire with two response options for each question: true or false. Despite a lack of agreement regarding its dimensionality and validity, it has been the most frequently used measure for the assessment of social desirability bias, having already been used in over one thousand studies and dissertations (Beretvas, Meyers, & Leite, 2002). Scores over 19 are

suggestive of high concern with social approval and a tendency to use self-presentation strategies. The full scale is presented in Appendix C.

Implicit attitudes related to personal morality were assessed by means of two separate *Implicit Relational Assessment Procedures* (Barnes-Holmes, Murphy, et al., 2010; Barnes-Holmes et al., 2006). The stimulus set for the Conceptual Morality IRAP ("CM-IRAP") was as follows:

- For label 1 ("I think it is good to"), consistent targets were: "Tell the truth", "Be fair", "Be honest", "Be moral", "Obey the law", "Behave well".
- For label 2 ("I think it is bad to"), consistent targets were "Tell lies", "Cheat", "Misbehave", "Be immoral", "Break the law", "Be dishonest".

The stimulus set for the Deictic Morality IRAP ("DM-IRAP") was as follows:

- Label set 1 consisted of the following stimuli: "I tell the truth", "I am good", "I am ethical", "I am honest", "I behave well", "I am fair". Consistent targets for this set were "Frequently", "Most often", "Very often".
- Label set 2 consisted of the following stimuli: "I am bad", "I deceive", "I defraud", "I lie", "I steal", "I cheat". Consistent targets for this set were "Rarely", "Seldomly", "Never".

Procedure. At the start of each experimental session the researcher thanked the participants for volunteering and attending, and informed them that they were about to take part in a preliminary study, designed to evaluate different types of tests in order to select the most appropriate ones for a forthcoming study. They were instructed to read the informed consent form (Appendix E), and reassured that their participation was voluntary, confidential, anonymous and that they had the right to terminate their participation at any

time, which none of them decided to do. After signing the informed consent, they were given €5 for their participation and were told that the money was theirs to keep independently of their performance or the exertion of their right to terminate their participation before completion of all the tasks. Participants were then presented with the cheating task, which was always given first because a pilot phase in which the IRAPs were presented first revealed that participants were much more likely to guess the true nature of the MMT.

The MMT. The task started with instructions presented on screen, which participants were asked to read carefully before proceeding with the equations. The instructions block read:

> "The following task is a classic cognitive experiment that looks at mathematical skills. You will be presented with an equation that we would like you to solve. It consists of 10 numbers between 1 and 20 that are to be added or subtracted. The equation will appear on screen, and then you MUST hit the SPACEBAR for the response box to appear. Please calculate the correct response as fast as you can, and enter it into the response box. Hit enter when you are done. You will be informed whether your answer is correct or incorrect. If it is incorrect, you will be given the opportunity to respond again until your answer is correct. Shortly after that, the next equation will appear. Altogether there are 20 such equations, presented in two blocks of ten equations each."

After reading the instructions, participants were told that once they pressed the Spacebar they had as much time as they needed to complete the equation, but that it was important to press it because otherwise this particular version of the program, allegedly unfinished, displayed the correct answer after some time. The researcher aimed to deliver this information in a neutral manner so as to not attract attention to the opportunity to cheat,

but at the same time making sure that participants understood that it was present. Having said this and made sure that the participant understood the instructions, the researcher then left the room to let the participant complete the MMT. After its completion, participants rated it on a scale from 1 (very boring) to 5 (very interesting), and proceeded to complete the IRAPs.

The IRAPs. We used the 2010 version of the IRAP software, at the time available for download at http://irapresearch.org/wp/downloads-and-training, on modern PCs (Dell™ OptiPlex™ 790) with standard 14" screens and running Windows 7. The software controls the presentation of instructions and stimuli, records participants' responses and outputs raw trial-by-trial data for analysis.

All IRAP tasks consisted of six blocks of 24 trials. Participants were first exposed to pairs of practice blocks, and were required to achieve two performance mastery criteria (detailed below) in order to proceed to a fixed set of three pairs of test blocks. Each trial started with two response options at the bottom of the screen, followed after 400 milliseconds by a label at the top and a target in the middle of the screen. There were four trial-types defined by combinations of the labels and targets according to a 2x2 cross-over of the label and target stimuli (see Figure 2). The program presented these four trial-types in a quasi-random order ensuring that each trial-type was presented six times within each block. The program also ensured that each trial-type was presented once in each sequence of four trials and that the same trial-type was never repeated across successive trials. Thus, in the Conceptual Morality IRAPs, the first trial-type involved presenting the label "I think it is good to" with a positive target such as "behave well". Another trial-type involved presenting the same label, but with a negative target ("Cheat"). The other two trial types presented the negative labels ("I think it is bad to") with the positive and negative targets, respectively.

These four trial-types are summarized here as Good/Moral, Good/Immoral, Bad/Moral, and Bad/Immoral (displayed in Figure 2).

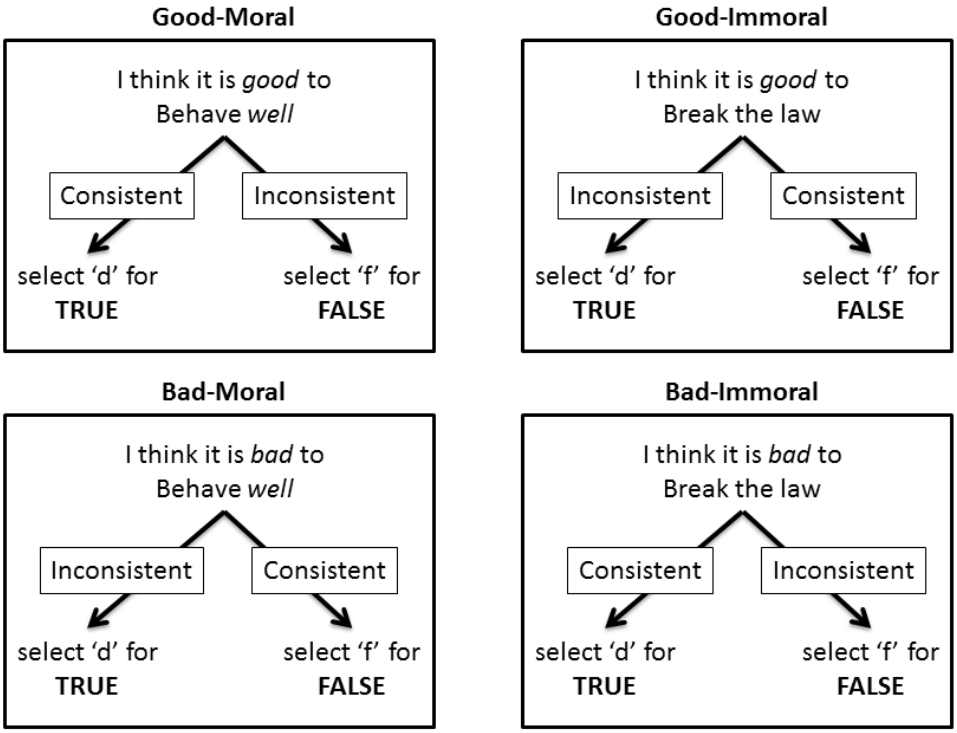


Figure 2. Visual representation of the four trial types in the CM-IRAP. Note that neither the words "consistent" and "inconsistent", nor the arrows, appeared on screen.

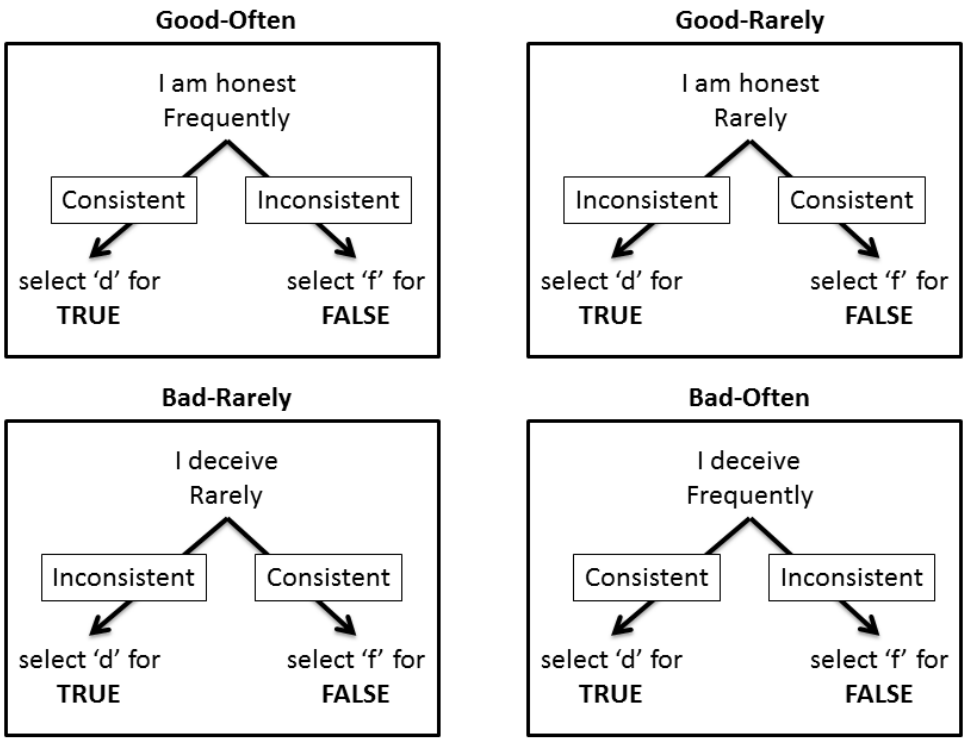


Figure 3. Visual representation of the four trial types in the DM-IRAP. Note that neither the words "consistent" and "inconsistent", nor the arrows, appeared on screen.

Pressing the key corresponding to the response deemed correct in each trial cleared the label and target stimuli from the screen, and 400 milliseconds later the next trial was presented. If participants emitted an incorrect response, a red X appeared immediately below the target and remained on screen until the correct response was emitted. If a participant failed to emit a response within 2500 milliseconds on each trial, the words "Too Slow" appeared directly below where the red X appeared and remained on screen until a response (correct or incorrect) was emitted.

At the beginning of each practice block, a message appeared on screen informing participants that it was a practice phase and that a few errors were expected, but also to try to avoid the red X (i.e., incorrect responses). The second block in every pair, both during practice and test phases, started with a message indicating that the previously correct and

incorrect responses were reversed, such that all odd numbered blocks required responding in a manner deemed consistent with common verbal practices – responding "True" on the first and last trial-types (Good/Moral and Bad/Immoral), and "False" on the other trial-types. During all even numbered blocks the opposite response was required, such that responses deemed inconsistent with common verbal practices were deemed correct (e.g., responding "False" on the Good/Moral trial-type). Responses deemed consistent versus inconsistent with common verbal practices are indicated in Figures 2 and 3, although it is important to note that these labels did not appear in the actual IRAP programme (i.e., participants were not told what was deemed consistent or inconsistent by the IRAP program or the researcher). Finally, at the end of each block, performance feedback appeared, which indicated the percentage of correct responses and median response latency for that block.

If the participant did not meet the accuracy and latency criteria in a pair of practice blocks, feedback for both blocks was presented along with a message that indicated that the participant was doing well but that it was a difficult task and they were invited to try again to reach the performance criteria. Once the criteria were met, a message appeared at the beginning of the next block and every block thereafter informing them that it was a test block. Participants were told to respond accurately and fast, trying to make as few mistakes as possible. The program progressed through the test blocks until all six were complete and then a blue screed appeared asking the participant to report to the researcher (who was waiting outside the experimental room).

Common configuration options for both IRAPs were the following: i) response options were labelled "True" and "False"; ii) response options did not appear in the same left-right position across more than three consecutive trials; iii) pairs of practice blocks were to be presented until participants achieved 80% correct responses and mean latencies under 2500 milliseconds in each block before proceeding to a fixed set of 6 test blocks. Performance

criteria were not required to proceed through the test blocks but accuracy and latency feedback were presented at the end of each block to encourage participants to maintain the performance criteria achieved during the practice blocks.

Instructions to participants. The two IRAPs were presented consecutively but in a counterbalanced order and started with a series of instructions based on the guidelines specified on the IRAP's "Experimenter's Script" (available for download at the website mentioned above). After sitting down and facing the computer, participants were told that the task comprised a number of trials and that their goal was to respond rapidly and accurately on each trial. They were informed that each trial displayed part of a sentence at the top of the screen and the rest in the centre, along with two response options at the bottom, which were always "True" and "False" but changed left-right positions across trials.

The response requirements were explained to the participants by pointing out that they could use the 'D' key to select the response option on the left side and the 'K' key to select the right-side option, and they were told to keep their index fingers resting lightly on those keys throughout each block of trials. They were informed that in each trial, they were to press the key that corresponded to the appropriate response option, and that they would receive feedback as to what constituted correct and incorrect answers – the latter being signalled by the appearance of a red X which would disappear when the correct key was pressed.

For the first two practice blocks the researcher sat beside the participant and instructed the individual on how to respond correctly and incorrectly on each trial. The experimenter focused on accuracy during the first two pairs of practice blocks, and then emphasized increasing speed during subsequent practice blocks. The experimenter also emphasized that the pattern of correct/incorrect responding would switch from block to block, so that participants needed to reverse their patterns of responding across blocks.

Participants were informed that the task was not asking them to express a particular opinion or belief but simply required them to respond as accurately and rapidly as possible across *all* blocks of trials. Once participants met the accuracy and latency criteria for a pair of practice blocks, the researcher withdrew from the experimental room stating that he would be just outside should the participant encounter any difficulty or wish to withdraw from the study. Only one participant communicated with the researcher before the IRAPs were completed and on this occasion he expressed a wish to terminate participation – he was thanked and debriefed immediately and his data were excluded from analysis.

Scales and Debriefing. Having finished the IRAPs, participants took a short two-minute break and were then asked to complete the scales, this time according to their own opinions. Finally, they were asked whether they saw something unusual or strange about the math task (as done by Von Hippel et al., 2005), told to read the disclosure information sheet presented in Appendix F. Finally, they were asked if they understood the reason for the temporary deception and the true nature of the study and whether they wanted the researcher to keep their responses for analysis, which they all agreed to. The procedure strictly adhered to the Guidelines for Safe Work Practice of the Department of Psychology and to relevant ethical guidelines as explained in Chapter 3.

Results and Discussion

Cheating measure. The software outputs a text file containing information about each trial: total time to give an answer, time before pressing the Spacebar to make the response box appear, response given, and, critically, whether the correct response was shown on the screen. Hence, cheating can be reported in absolute terms (cheaters vs. non-cheaters) and also in terms of magnitude (number of cheats). None of the participants reported guessing the true goal of the task after its completion. Each of the two blocks consisted of ten trials: if the participant did not press the Spacebar during the Slow block the

answer appeared after 10 seconds from the presentation of the equation, and in the Fast block the answer appeared after 1 second. Table 1 shows the main descriptive statistics for the cheating measure.

Table 1

Descriptive Statistics for the MMT in Study 1

Task	% who cheated	Average cheats	*SD*	Range
Slow block only	35.1	0.59	0.95	0-3
Fast block only	64.9	1.57	1.62	0-6
Combined	67.6	2.16	2.23	0-7

A majority of the sample waited to see the answer at least once, which is in line with findings in other studies. Von Hippel et al. (2005) found that 79-85% across different experiments with the task waited to see the answer (i.e., "cheated"), with a slightly wider range (0-10) and comparatively more cheating in the second block (fast task), which we also observed. Jordan et al. (2011) found that 57% of the sample allowed the answer to appear at least once. However, a study by Teper and Inzlicht (2010) reported that most participants did not cheat at all, even with a monetary payoff, which probably calls attention to a potential susceptibility of the task to external factors. We found no significant correlation between the rating given to the task (in terms of boredom) and the amount of cheating.

Scales. Consistent with Caprara et al. (2009), and as expected, the two moral disengagement scales correlated highly with each other ($r = 0.79$, $p < 0.01$). Table 2 shows descriptive analyses for the moral disengagement and social desirability scales.

Table 2.

Descriptive Statistics for the Scales in Study 1

Scale	Range	Mean	SD
Civic Moral Disengagement (CMD)	46-107	69.81	13.49
Moral Disengagement (MDS)	4-36	15.08	7.52
Social Desirability Scale (MC-SDS)	7-29	15.95	5.35

IRAPs.

Data preparation. Response latencies, or the elapsed time in milliseconds between the onset of the trial and the emission of a correct response by the participant, are the primary datum from the IRAP. The presentation software outputs raw trial-by-trial latency data and calculates scores for analysis according to an algorithm derived from the standard transformation of IAT scores (H. Cai, Sriram, Greenwald, & McFarland, 2004). This procedure has been shown to minimise the impact of individual differences in age, IQ and speed of responding, amongst other individual factors (Barnes-Holmes, Murtagh, et al., 2010; Power, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009), and specifically involves the following steps:

(a) remove the entire data set with more than 10% latencies below 300 milliseconds.
(b) remove latencies at or above 10.000 milliseconds from the dataset;
(c) calculate the standard deviation across each pair of test blocks for each of the four trial types, yielding 12 standard deviations;
(d) calculate mean latencies for each trial type within each block of test trials, yielding 24 means;
(e) subtract the mean latency for each trial-type in test block 1 from test block 2, and repeat for test blocks 3 and 4, and for test blocks 5 and 6, to obtain twelve difference scores;
(f) divide the difference scores obtained in the previous step by their associated standard deviations obtained in step c;

(g) obtain four mean *D*-IRAP scores, one for each trial type, by averaging the scores from the three pairs of test blocks.

Only test block data from participants who completed both IRAPs were used for this calculation. The data for nine participants were excluded from the analyses because they failed to meet the performance criteria during practice blocks in either of the IRAPs, and therefore never progressed to the test blocks.

Moral biases. Mean *D*-IRAP scores for each trial type, obtained from step g above, can be plotted into a figure, and visual inspection of the direction and height of the bars can be used as a rough estimation of pro- or anti-moral bias: bar height indicates the strength of the effect and the direction (positive or negative) indicates the nature of the bias. In the case of the CM-IRAP, the overall mean *D*-IRAP scores for the four trial-types presented in Figure 4 indicate that, in general, participants responded "True" more quickly than "False" on the *Good/Moral* and *Bad/Immoral* trial-types, and responded "False" more quickly than "True" on the *Bad/Moral* trial-type. The *D*-IRAP effect was strongest for the Good/Moral trial-type in both groups and weakest, in fact approaching zero, for the Good/Immoral trial-type. Four one-sample t-tests indicated that the IRAP effect was significantly different from zero for the Good/Moral ($t = 5.814$, $p < 0.01$) and Bad/Moral ($t = 2.273$, $p < 0.05$) trial-types only.

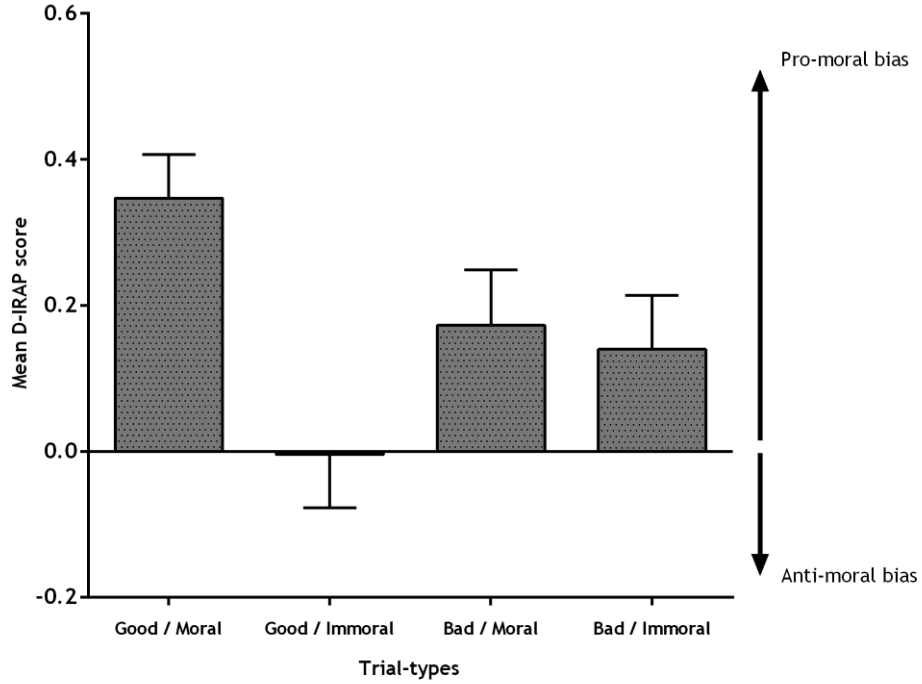


Figure 4. Mean *D*-IRAP scores with standard error bars for the four trial types of the conceptual (CM) IRAP. Positive numbers indicate a pro-moral bias, and negative numbers indicate an anti-moral bias.

The mean *D*-IRAP scores for the DM-IRAP are shown in Figure 5. The scores indicate that in general participants responded "True" more quickly than "False" on *Good/Often* and *Bad/Often* with relatively weak effects observed for the remaining two trial-types. One-sample t tests revealed effects that the effects for Good/Often ($t = 5.779$, $p < 0.01$) and Bad/Often ($t = -0.370$, $p < 0.01$) were indeed significantly different from zero, but the effects for the other two trial-types were not.

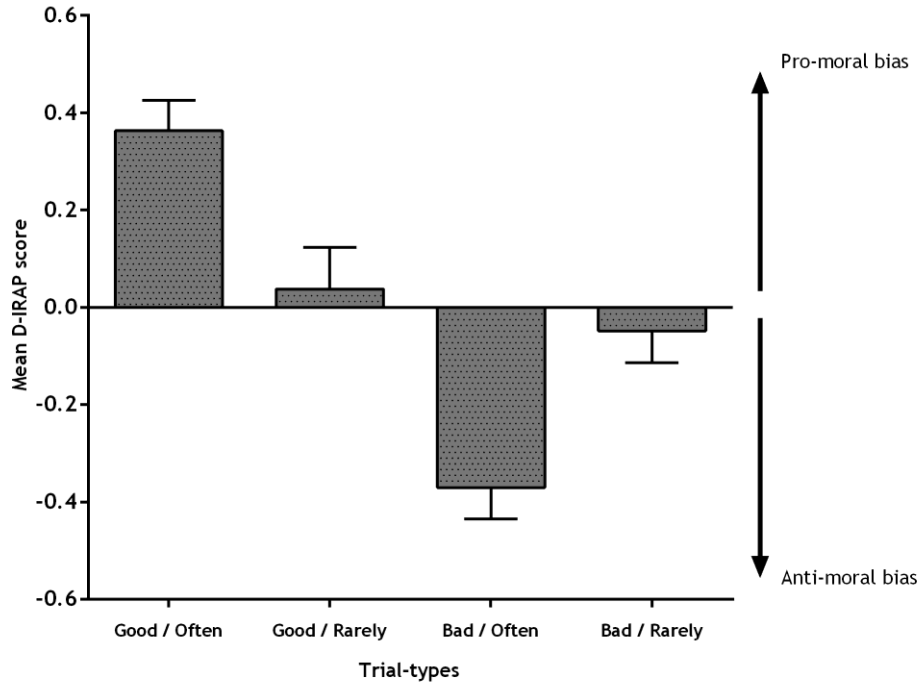


Figure 5. Mean *D*-IRAP scores with standard error bars for the four trial types of the deictic (DM) IRAP. Positive numbers indicate a pro-moral bias, and negative numbers indicate an anti-moral bias.

Predicting cheating behaviours. With the goal of determining whether the IRAPs could predict immoral behaviour, Pearson correlations were calculated between the total MMT score and the mean *D*-IRAP scores for the four trial types in each IRAP (see Table 3). Of the eight correlations, two proved to be significant, one from the CM-IRAP for the *Bad/Immoral* trial-type and one from the DM-IRAP for the *Bad/Often* trial-type. The former correlation indicates that an increasing bias towards confirming that bad actions are immoral predicts more cheating responses on the MMT. The latter correlation indicates that a bias towards denying a high frequency of immoral behaviour also predicts more cheating on the MMT.

Table 3.

Correlations between MMT score and Mean D-IRAP Scores for both CM- and DM-IRAPs

Trial type	Correlation with MMT	
	r	*p*
CM-IRAP		
Good/Moral	-0.147	> 0.3
Bad/Moral	-0.109	> 0.4
Good/Immoral	-0.037	> 0.5
Bad/Immoral	0.428*	0.02
DM-IRAP		
Good/Often	0.072	> 0.5
Good/Rarely	0.111	> 0.4
Bad/Often	0.380*	0.04
Bad/Rarely	0.162	> 0.3

(*) Significant at the $p < 0.05$ level; (**) Significant at the $p < 0.01$ level.

Moral disengagement. No significant correlations were found between the total scores in the moral disengagement scales and the mean *D*-IRAP scores.

Split-half correlations. Following other studies, overall split-half reliability scores were calculated for each IRAP in order to provide a measure of internal consistency, and this yielded a moderate and significant result for the DM-IRAP ($r = 0.48$, $p < 0.01$), but not the CM-IRAP.

Summary and Conclusions. Overall, the results of this first study for the CM-IRAP produced three IRAP effects that were consistent with common sense, in that participants confirmed *Good/Moral* and *Bad/Immoral* relations more quickly than they denied them, and

denied *Bad/Moral* relations more quickly than they confirmed them. However, there was no clear bias when responding to *Good/Immoral* relations on the IRAP. The DM-IRAP yielded results that were even more counter-intuitive. First, participants revealed a tendency to confirm that they often engaged in both good and bad behaviours. Second, participants failed to show any clear biases on the remaining two trial-types, which required responses to questions concerning how rarely they engaged in good or bad behaviours. Although the emergence of putatively counter-intuitive results raise important questions, which we will address later, it seemed important to replicate these effects before drawing any strong conclusions.

Another possibly counter-intuitive result that emerged was the finding that increased bias scores in a pro-moral direction for the Bad/Immoral trial-type (in the CM-IRAP) and the Bad/Often trial-type (in the DM-IRAP) appeared to predict increased cheating on the MMT. In other words, it appears that participants who more strongly confirmed that bad actions are immoral, and denied engaging in immoral behaviour, tended to cheat more. One possible explanation for this finding might be that individuals who tend to cheat, lie and deceive may well be more practiced at criticising immoral actions and denying that they engage in such behaviours precisely because doing so is an example of such behaviour itself. Insofar as the IRAP is a measure of the relative probabilities of specific verbal relations, then these correlations may not be so counter-intuitive.

Once again, it is important to recognise that the current findings were obtained from only one exploratory study and thus it was deemed important to attempt to replicate the results in a second study before speculating too broadly. With this in mind, the second study reported in the current thesis employed the same two IRAPs employed in Study 1 but adopted a different cheating task. A different cheating task was employed at this point in the research programme due to ad hoc comments that were provided by some participants in

Experiment 1. Specifically, they reported that the answer appeared too quickly in the second block and they were focused on reading the equation and temporarily forgot to press the Spacebar. Consequently, it is possible that at least in some cases, the participant did not intend to see the answer, but was distracted and failed to prevent it from showing up.

Study 2: Replicating Study 1 Using an Alternative Cheating Measure

Even though the MMT had been used as a way to operationalise immoral behaviour in more than twenty studies at the time of the current research was being conducted, as noted above there was a potential that false-positive cheating responses may have been recorded. In Study 2, therefore, we explored the idea of creating a different cheating measure based on the dice-in-a-cup task by Fischbacher and Heusi (2008). The original task involves participants privately rolling a die under a cup, memorising the number and then completing a purportedly unrelated questionnaire, after which they get monetary compensation depending on the reported result of the die roll. The distribution of reported rolls is typically compared to the expected uniform distribution, and if numbers below 4 are significantly less reported, this suggests that many participants lied about the number they actually rolled.

The original task is inadequate for our purposes because it makes it impossible to observe deception at an individual level, a requirement for our research programme given that we aimed to relate individual performances on the IRAP with the cheating task. Indeed, the problem of identifying cheating at the individual level is a limitation of other cheating tasks that have also been used to operationalise immoral behaviour, and they are hence unsuitable as well. Therefore, we decided to create a computerised task ("Dice Cheating Task" or DCT), in principle would provide a measure of individual cheating responses. We also used all the other measures from Study 1 except for the MMT.

Materials and Methods

Participants. A sample consisting of 30 undergraduate students from the National University of Ireland, Maynooth (18 females, 60%) aged 18 to 25 (M = 19.23, SD = 1.63) contributed to this study. Participants were contacted from a volunteer pool assembled through announcements made at the start of each academic year. Upon their arrival at the

laboratory, they were again told that the researcher intended to get feedback on a number of tasks in order to select the appropriate ones for a forthcoming study. Next, they read and signed a standard informed consent form (Appendix E), although the justified deception from Study 1 was maintained due to potential reactivity. Immediately after signing the consent form and having been given the opportunity to ask any questions about the procedure, they were given €5 for their participation. Two participants failed to achieve criteria to proceed to the test blocks in one of the IRAPs and their data were excluded from analysis, leaving us with a sample of 28 participants.

Measures. Study 2 tested a new cheating task and aimed to replicate the results from Study 1 thus providing convergent validity. The same scales and IRAPs from the previous study were used, but cheating was operationalised in terms of performance in a Die Cheating Task (DCT). The task is presented to participants as an attentional task with the goal of getting a high score. Participants are presented with a computer interface similar to that shown in Figure 6.

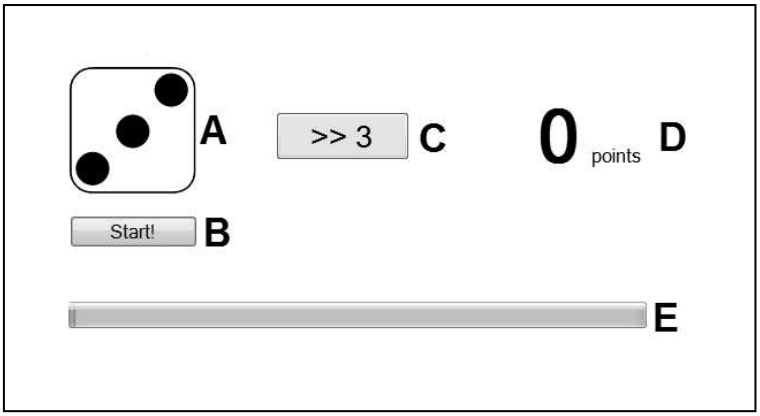


Figure 6. Main Interface of the Die Cheating Task, showing the five main elements: A) a simulated die; B) a start/stop button that initiated or stopped the simulated roll of the die; C) a button that displayed the number rolled and that had to be clicked in order to increase the score; D) a score counter, and E) a progress bar that served as a visual indicator of the remaining number of trials (Letters added to the figure for clarity).

The task, coded in Visual Basic Express 2012 specifically for this experiment, was presented on the same type of computer as Study 1 and in the same location. At the beginning of the task, a tutorial was presented consisting of a set of PowerPoint slides that explained how to interact with the task. The researcher sat beside the participant during this presentation to ensure that the instructions were being understood and to answer any questions that might arise. The tutorial (slides reproduced in Appendix G) informed participants that the task tested their reflexes and attention and that their goal was to get a score as high as possible by adding the results of simulated dice rolls to a tally counter. Critically, participants were required to stop each roll of the die and thus getting a high score involved stopping the roll when the die was displaying high numbers (i.e., 5 or 6). This task was complicated by the fact that the roll was fast and stopping it at the desired number required them to be very focused and attentive.

The tutorial also informed that they could interact with the task by pressing the "Start/Stop" button (B in *Figure 6*) to roll the die or the "Add to score" button (C in the same figure) to increase their score by the number displayed on the face of the dice. When the "Start/Stop" button was pressed, the roll was simulated by replacing the image of the die with another randomly selected face of the die followed by a blank face after 220 milliseconds. This produced the effect of a rapidly changing die face. Pilot work showed that participants could clearly see the numbers, but that pressing the Stop button when a desirable (i.e., high) number was being displayed was challenging given the short time that the number remained on screen.

The task involved 80 trials. 50 of those were "normal" trials in which the "Add to score" button displayed the same number that the die was showing when the roll stopped. However, in 30 specific trials the computer always displayed pre-selected combinations of die rolls and scores to be added to the tally counter. Figure 7 shows these special trials: on

High Roll-Low score (HL) trials, the die would show a high number (5 or 6) but the button would show a low number (2 or 3). On Low Roll-High Score (LH) trials, the opposite was true – the die would show a lower number than the button. Cheating was defined as adding scores in the LH trials, since the payoff in terms of score was higher. HL trials were not considered cheating and served as a way to determine if the participants had understood the tasks and were performing it according to the instructions. The tutorial clearly indicated that the program would sometimes display different numbers in the die and the button, in which case they were supposed to just roll the die again and not click the "Add to score" button.

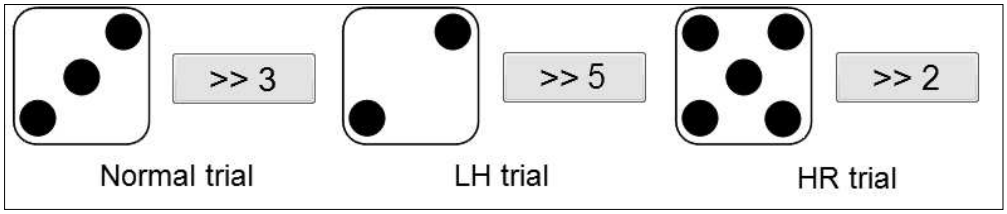


Figure 7. The three types of trials in the DCT, showing the relation between the number on the die and the number on the button. On normal trials, both numbers match. On LH trials, the die shows a lower number than the button, and in HR trials, the die shows a higher number than the button.

The magnitude of the difference between the number rolled and the number displayed by the button can be used to further characterise the cheating (i.e. LH) trials: adding a 4 when the roll is actually 3, for example, is a "small" cheat. But adding a 5 when the roll is actually a 2 involves aiming for a bigger payoff. We hypothesised that people were more likely to engage in cheating when the distance between the expected and deceptive behaviours was smaller, which is consistent with previous research (Hilbig & Hessler, 2012).

After 80 trials, the task asked participants to restate the nature and the goal of the task and to give two ratings from 0 to 10: how difficult and how entertaining the task was. A preliminary pilot phase (n = 15) using the DCT was conducted to test its properties, without using the other measures. All the participants in the pilot phase reported at the end that they

believed that the task was testing their reflexes and their attention and during debriefing none of them reported thinking that it was a measure of deception, although a few of them (n=6, 40%) revealed that they had taken advantage of the LH trials to increase their score (i.e., they cheated).

Insofar as this study intended to replicate Study 1 with a different deception measure, the same instruments from Study 1 were used, namely the two IRAPs (Conceptual and Deictic), the moral disengagement scales and the social desirability scale.

Procedure. The experiment took place in one of the experimental cubicles at the Department of Psychology, where participants were given similar instructions to those from Study 1, specifically that their performance on different types of tests was going to be used to select the most appropriate ones for future research. This was followed by reading and signing the consent form, reassurance of all their rights as participants and the delivery of a monetary compensation of €5.

Completion of these preliminary procedures was followed by the DCT. The researcher started the tutorial, stayed in the room with the participant to ensure that the instructions were read and understood, answered whatever procedural questions arose and started the actual task and left the experimental room to let the participant complete the task.

Upon conclusion of the DCT, the session proceeded exactly as in Study 1 at this point, in that participants were given the two IRAPs in a counterbalanced fashion and then they completed the questionnaire measures. Finally, they were thoroughly debriefed in the same manner described in Study 1, thanked for their participation and asked if they wanted their data to be kept or erased. All the participants consented to having their data used for the study.

Results and Discussion

Data analysis followed the same principles detailed for Study 1, except for the cheating measure. In this case, the presentation software saved a file containing all the relevant information for each trial, and specifically whether the participant availed of the cheating mechanism.

Cheating measure. 46.4% of the sample increased their score by clicking the "Add to score" button when the die displayed a lower number, which was our operational definition of cheating. None of the participants clicked the button when a lower number was being displayed though, which suggests that all the participants were sufficiently engaged in the task as to ignore the High roll-Low score trials. All the participants correctly responded that the goal was to get the highest score when asked at the end of the task. As predicted, 42.8% of the sample engaged in "small" cheats but only 21.4% in "big" cheats, the difference between the two variables being statistically significant in a paired samples t-test ($t = 2.357$, $p < 0.03$). The average rating of the task on a scale from 1 (boring) to 10 (entertaining) was 7.8, and in terms of difficulty the average rating was 5.7 / 10.

Scales. Again, and as expected, the moral disengagement scales (CMD and MDS) showed a moderately high correlation ($r = 0.53$, $p < 0.05$). No correlation was found between the cheating measure and the overall moral disengagement scores in either of the two scales. Table 4 presents descriptive statistics for the scales.

Table 4.

Descriptive Statistics for scales in Study 2

Scale	Range	Mean	SD
Civic Moral Disengagement (CMD)	53-101	80.47	10.35
Moral Disengagement (MDS)	1-27	14.40	5.47
Social Desirability Scale (MC-SDS)	5-26	14.43	4.31

IRAPs. Overall mean *D*-IRAP scores for the four trial-types in the CM-IRAP are presented in Figure 8. In general, participants responded "True" more quickly than "False" on the *Good/Moral* and *Bad/Immoral* trial types. This pattern of responding was also observed, albeit on a smaller scale, in the *Good/Immoral* trial-type, but was reversed in the *Bad/Moral* trial-type. In general, the distribution of effects resembled that of the CM-IRAP in Study 1. To evaluate if the mean scores were significantly different from zero, we performed four one-sample t-tests that indicated a statistically significant effect on the *Good/Moral* (t = 3.473, $p < 0.01$) and the *Bad/Immoral* (t = 3.341, $p < 0.01$) trial-types only, suggesting once again that participants readily responded to good actions as moral and bad actions as immoral.

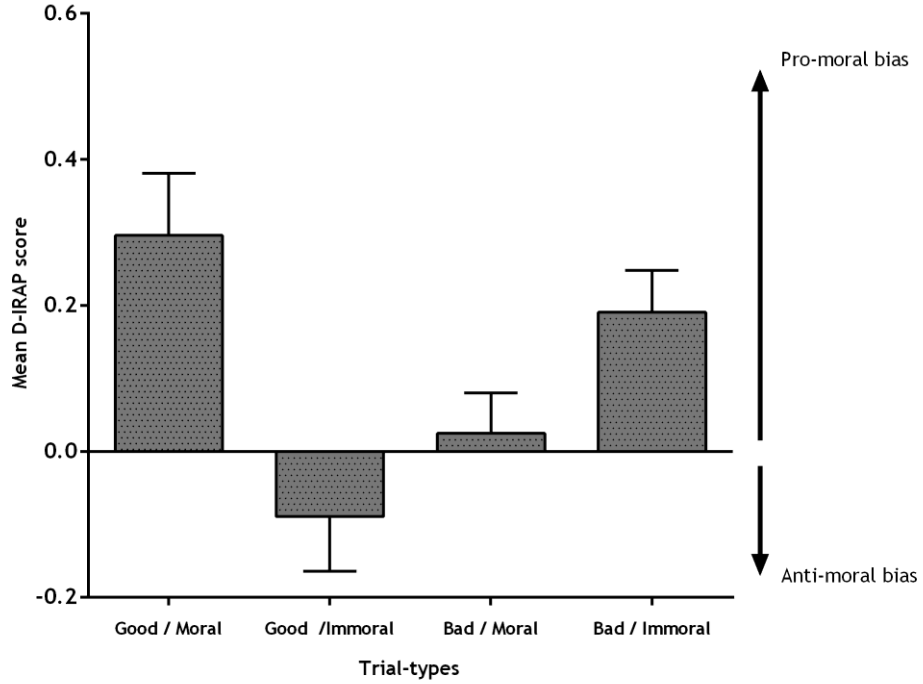


Figure 8. Mean D-IRAP scores with standard error bars for the four trial types of the conceptual (CM) IRAP. Positive numbers indicate a pro-moral bias, and negative numbers indicate an anti-moral bias.

The DM-IRAP is presented in Figure 9. The effects displayed in this IRAP are not unlike those of Experiment 1, especially in the trial types where targets correspond to the "Often" category. The strongest effects were for the *Good/Often* and *Good/Rarely* trial types and both showed pro-moral biases. One-sample t tests indicate that the mean scores were significantly different from zero for *Good/Often* ($t = 7.783$, $p < 0.01$) and for *Good/Rarely* ($t = 3.014$, $p < 0.01$). The effect for the *Bad/Often* trial-type was significant in Study 1 and approached significance in the current study ($t = -1.761$, $p = 0.08$); again the bias was in an anti-moral direction.

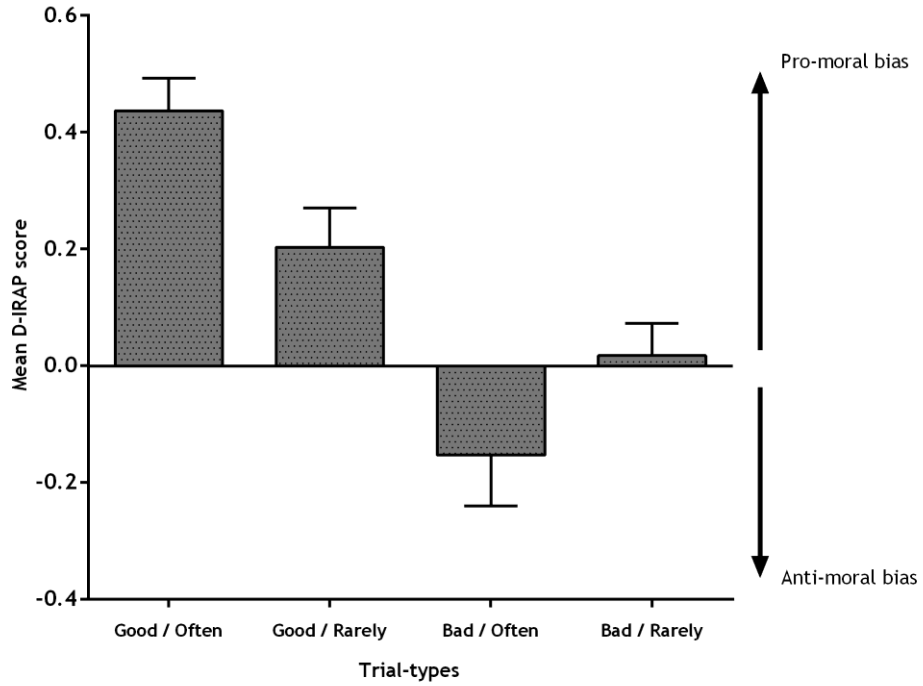


Figure 9. Mean D-IRAP scores with standard error bars for the four trial types of the conceptual (DM) IRAP. Positive numbers indicate a pro-moral bias, and negative numbers indicate an anti-moral bias.

Predicting cheating. We calculated Pearson correlations between the three different scores offered by the DCT ("small" cheats, "big" cheats and total cheats) and the mean *D*-IRAP scores for the four trial types in each IRAP. The resulting 2x4x3 correlation matrix is presented in Table 5. Even though there were no statistically significant correlations with an $\alpha < 0.05$, two trial-types in the DM-IRAP, namely *Good/Rarely* and *Bad/Often,* showed weak to moderate correlation coefficients, which approached significance, with number of "small" cheats. The correlation for the latter trial-type (*Bad/Often*) was similar to that observed in Study 1, indicating that a bias towards denying engaging in bad behaviour predicted higher levels of cheating. The correlation for the *Good/Rarely* trial-type, which was specific to this study, indicates that a bias towards denying that one is rarely good predicts increased cheating responses. In both cases, therefore, the correlations on the trial-types that involved *denying* immoral behaviour appeared to predict cheating. This outcome is consistent with the

argument that denying dishonesty on the IRAP may be reflective of well-practiced verbal responses for individuals who engage in relatively high levels of cheating in the natural environment (i.e., denying that they are "cheats"). A correlation that approached significance was also recorded for the CM-IRAP between the *Bad/Moral* trial-type and big cheats, but the N was so low (6 participants) interpreting this result would be unwise.

Table 5.

Correlations between DCT scores and Mean D-IRAP Scores for both CM- and DM-IRAPs

	Small cheats		Big cheats		Total cheats	
Trial-types	*r*	*p*	*r*	*p*	*r*	*p*
CM-IRAP						
Good/Moral	0.07	> 0.5	-0.06	> 0.5	0.04	> 0.5
Bad/Moral	0.04	> 0.4	0.32	**0.08**	0.16	> 0.3
Good/Immoral	0.06	> 0.5	0.05	> 0.5	0.03	> 0.5
Bad/Immoral	-0.21	> 0.1	-0.08	> 0.5	-0.18	> 0.2
DM-IRAP						
Good/Often	0.06	> 0.5	0.24	> 0.1	0.15	> 0.3
Good/Rarely	0.33	**0.08**	0.12	> 0.4	0.28	> 0.1
Bad/Often	0.34	**0.07**	0.05	> 0.6	0.28	> 0.1
Bad/Rarely	0.27	> 0.09	0.15	> 0.3	0.24	> 0.1

Moral disengagement. Once again, no significant correlations were found between the total scores from the moral disengagement scales and the mean *D*-IRAP scores.

Summary and Conclusions. The IRAP effects observed here resemble the ones from Study 1 in many ways. The CM-IRAP produced effects that are consistent with intuitive, common sense expectations of faster confirmation of *Good/Moral* and *Bad/Immoral* relations.

Once again, however, no clear bias could be detected when responding to *Good/Immoral* or *Bad/Moral* relations. In the DM-IRAP in the previous study, participants revealed a counter-intuitive bias towards confirming that they engaged in bad behaviours often – this effect was also present in the current study, even though this trend did not reach statistical significance.

At this point in the research programme we had employed two different IRAPs (CM and DM) with two different cheating tasks (the MMT and DCT). In Study 1 we found significant correlations between one trial-type in each IRAP (CM-IRAP, *Bad/Immoral*; DM-IRAP, *Bad/Often*) and the cheating measure. In Study 2 we failed to record any significant correlations, although two of them approached significance and were generally consistent with the interpretation of the results we offered previously – that people who tend to cheat in the natural environment will be more highly practiced at denying that they do and this may be reflected on IRAP trial-types that target "denial-based" verbal responses.

Although there was some overlap in the results from Study 1 to Study 2, a number of concerns arose at this point. The first is that the DCT failed to generate the same variance in cheating behaviour relative to the MMT, thus potentially undermining the DCT's usefulness as a laboratory measure of cheating behaviour. A directly related concern was that the correlations between the trial-types and the cheating measure in the second study only approached significance, although the trends from Study 1 were still observed. The remaining studies reported in the current thesis therefore employed a slightly modified version of the MMT, rather than the DCT.

The third concern was that, once again, there was no correlation between the moral disengagement scales and the cheating measure or the IRAPs in both studies. As noted previously, it is possible that the self-report scales were subject to self-presentation effects and thus at this point, given the lack of correlations, we ceased using them for the remainder of the current research programme. The final concern that emerged following Study 2 was

recognition that the DM-IRAP may have been targeting only a limited aspect of deictic relations, namely frequency of cheating. Much of the research on cheating, however, highlights the important role played by maintaining a moral sense of self even when engaging in immoral behaviour (Aquino & Reed, 2002; Bandura et al., 1996). To address all this concern, in Study 3 we employed an IRAP that was designed to target how participants feel when they engage in good and bad behaviours.

Chapter 5

Deictic Framing and Feelings as a key component of morality

Study 3. Exploring the role of Feelings in Deictic Responding

As explained previously, in Study 3 we decided to modify the DM-IRAP, such that it would target feelings associated with moral and immoral behaviour rather than simply frequency. In addition to the foregoing changes for Study 3, we also introduced a self-report instrument that was designed to target the construct of "psychopathy". We introduced this measure based on the argument, outlined previously, that a degree of verbal empathy might be needed to encourage individuals to follow moral rules. Specifically, based on an understanding of the consequences, especially the pain felt by others, through transformations of stimulus functions, engaging in behaviours that are coordinated with verbal labels such as "good", and "moral" may be inherently reinforcing. Indeed, in mainstream psychology, low levels of empathy and remorse, together with behavioural boldness, are all part of the construct of psychopathy (Scott, 2014). In fact, Sobhani and Bechara (2011) suggest that people who engage in immoral behaviour or corruption might have personality traits resembling those of psychopaths. With this in mind, and given our theoretical interest in these attributes, we decided to explore whether a well-known measure of psychopathy would correlate with the cheating measure (MMT) and one or more of the trial types in the IRAP.

Materials and Methods

Participants. 33 students from the National University of Ireland, Maynooth, aged between 18 and 25 (M = 19.03, SD = 1.26) took part in this study. They were recruited through classroom announcements during which they could register their interest in participating in psychological research. In the end, four participants did not achieve the criteria required to proceed to the test blocks in at least one of the IRAPs, and consequently their data were excluded from analysis, leaving a sample of 29 participants (62.1% females).

Measures. Our cheating measure for this study was a slightly revised version of the MMT. In the original, it could be argued that some of the participants were simply slow responders or were just distracted momentarily and failed to press the spacebar in time to prevent the answer from showing up on screen (i.e., a cheating response was registered but it should not have been recorded as such). A simple fix suggested by Jordan, Mullen and Murnighan (2011) was therefore implemented, namely that the answer that appeared was +/- 1 from the mathematically correct answer. The software signalled whether the participant entered the correct response or the altered one – the latter (incorrect value) was taken as a deliberate cheat, because it implied that the participant had not actually performed the calculation, but just entered the number that showed up on the screen, even though it was incorrect. The task was otherwise identical to the original.

Once again, we used two IRAPs in this study. The CM-IRAP was the same as in the previous experiments, but the new deictic IRAP referred to feelings when engaging in moral and immoral actions. This modified version, that we called the DF-IRAP, involved presenting two sets of label stimuli, the first of which was the phrase "When I" followed by "Behave well", "Play by the rules", "Obey the law", "Do the right thing", "Act morally", and "Tell the truth". The second set of label stimuli also presented the phrase "When I" but was followed by: "Break the law", "Cheat", "Act immorally", "Tell lies", "Break the rules", and "Deceive". The two types of target stimuli were the phrases "I feel good" and "I feel bad". The four trial-types for the DF-IRAP may thus be summarized as: 1. *Do-Good/Feel-Good;* 2. *Do-Good/Feel-Bad;* 3. *Do-Bad/Feel-Good; and* 4. *Do-Bad/Feel-Bad*, and are presented in Figure 10.

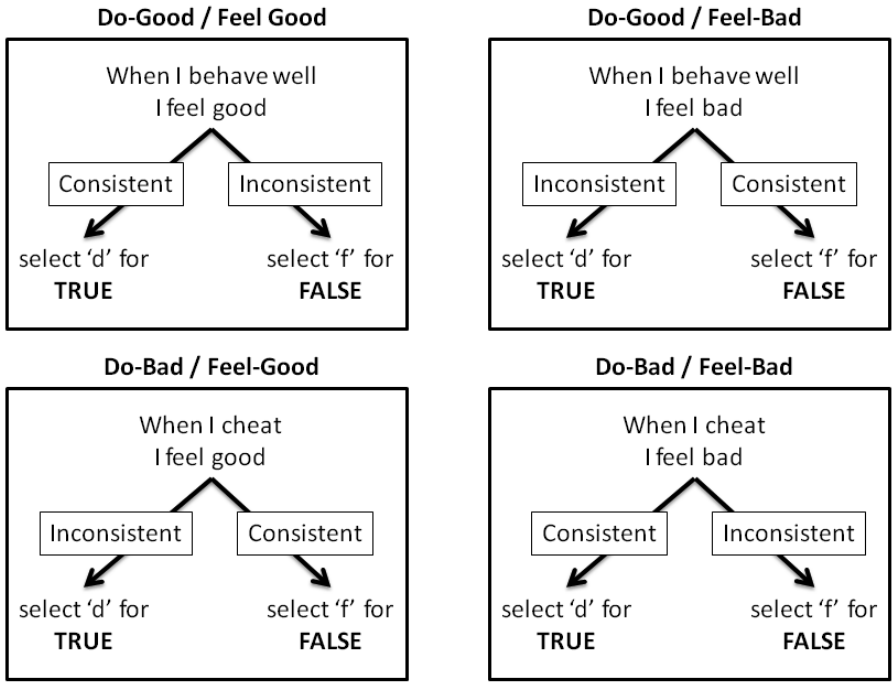


Figure 10. Visual representation of the four trial types in the DF-IRAP. Note that neither the words "consistent" and "inconsistent", nor the arrows, appeared on screen.

In this study we used the revised 2012 version of the IRAP software. This version is procedurally identical to the previous version, but it permits the researcher to insert a specific rule, which appears at the beginning of each practice and test block. The rules that were inserted in to the CM-IRAP were as follows: Consistent blocks, "*Please answer AS IF GOOD actions were MORAL and BAD actions were IMMORAL*"; Inconsistent blocks, "*Please answer AS IF GOOD actions were IMMORAL and BAD actions were MORAL*" The rules that were inserted into the DF-IRAP were as follows: Consistent blocks "*Please answer AS IF being GOOD makes you feel GOOD and being BAD makes you feel BAD*"; Inconsistent blocks, "*Please answer AS IF being GOOD makes you feel BAD and being BAD makes you feel GOOD*".

As a measure of Psychopathy we used the widely known *Levenson Psychopathy Scale* (LPS), self-report questionnaire consisting of 26 items (7 reversed to control for acquiescence). The LPS has the distinct advantage of being one of the few psychopathy

measures specifically intended for a non-clinical population, and it is based on the two-factor model of psychopathy, which proposes the existence of primary and secondary psychopathy. As such, 16 items of the scale evaluate primary psychopathy, or a general disposition towards callousness, manipulation, selfishness and lying. Examples of items for this factor are "*Looking out for myself is my top priority*" or "*I enjoy manipulating other people's feelings*". Secondary psychopathy, targeted by the remaining 10 items, is a contextually-mediated and situational engagement in antisocial and immoral behaviour, with negative emotions such as anxiety, fear and remorse that are not present in primary psychopathy (Dean et al., 2013). This factor is assessed in the scale with items such as "*When I get frustrated, I often 'let off steam' by blowing my top*" and "*Before I do anything, I carefully consider the possible consequences*". The item response format is a 5-point Likert scale from 1 to 5, which permits a range of 13 to 65 for each subscale, and 26 to 130 for the total scale. The reported Cronbach's alpha for the scale was 0.82. (Levenson, Kiehl, & Fitzpatrick, 1995). We tested the scale in a small sample and found problems with item 6 ("I let others worry about higher values; my main concern is with the bottom line"), which was replaced by "I let others worry about higher values; my main concern is with the bare necessities." (suggested by Hauck-Filho & Teixeira, 2014).

Procedure. Participants were invited to the Psychology Laboratory, thanked for volunteering and attending the session, and given similar instructions to those from the previous studies. Specifically, they were told that the study intended to evaluate different types of tests in order to select a few for future studies based on performance. They read and signed a standard consent form (Appendix E) and received reassurance of confidentiality, anonymity and right to withdraw from the study at any time. Then they received €5 for their participation and started the session by completing the revised MMT in the exact same manner as described in Study 1.

After completion of the cheating measure, they were presented with the two IRAPs in a counterbalanced order, such that one group started with the CM-IRAP and the other group started with the DF-IRAP. These were followed by administration of the Levenson Psychopathy Scale. At the end, they underwent full debriefing. None of the participants declined to have their data recorded and used for further analysis, and they all reported that they understood the reason for the deception involving the MMT.

Results and Discussion

Cheating measure. Cheating levels in the modified MMT were similar to those observed in previous studies and in Experiment 1 of the current thesis. Table 6 presents the percentage of participants who availed of the opportunity to cheat in each block, and the average number of cheating responses with standard deviations and ranges.

Table 6

Descriptive statistics for the MMT.

Task	% who cheated	Average cheats	*SD*	Range
Slow block only	58.6	1.03	1.08	0-3
Fast block only	58.6	1.41	1.70	0-7
Combined	68.9	2.22	1.24	0-8

Psychopathy. Table 7 presents descriptive statistics for the Levenson Psychopathy scale. There are no recommended cut-off points for the scale: Levenson et al. (1995) found a mean of 29.13 (*SD* = 6.86) for Primary Psychopathy and 19.32 (*SD* = 4.06) for Secondary Psychopathy in college students in the United States. More recently, Falkenbach, Poythress, Norman and Creevy (2008) found means of 31.93 (*SD* = 9.01) and 20.93 (*SD* = 4.99) in a similar sample of college students.

Table 7.

Descriptive statistics for the Levenson Psychopathy Scale.

Sub-scale	Mean	*SD*	Range
Primary Psychopathy	34.41	4.96	24-42
Secondary Psychopathy	21.41	2.36	15-26
Total score	55.82	6.25	44-65

IRAPs. The mean *D*-IRAP scores for the four trial types in the CM-IRAP are presented in Figure 11. Continuing with the trend of the previous studies, participants tended to respond "True" more quickly than "False" to the *Good/Moral* and *Bad/Immoral* trial types, and respond "False" more quickly than "True" on the *Bad/Moral* trial-type. The effect on the *Good/Immoral* trial-type was almost zero. One-sample t tests revealed a significant effect only for the *Good/Moral* trial type ($t = 5.56$, $p < 0.01$).

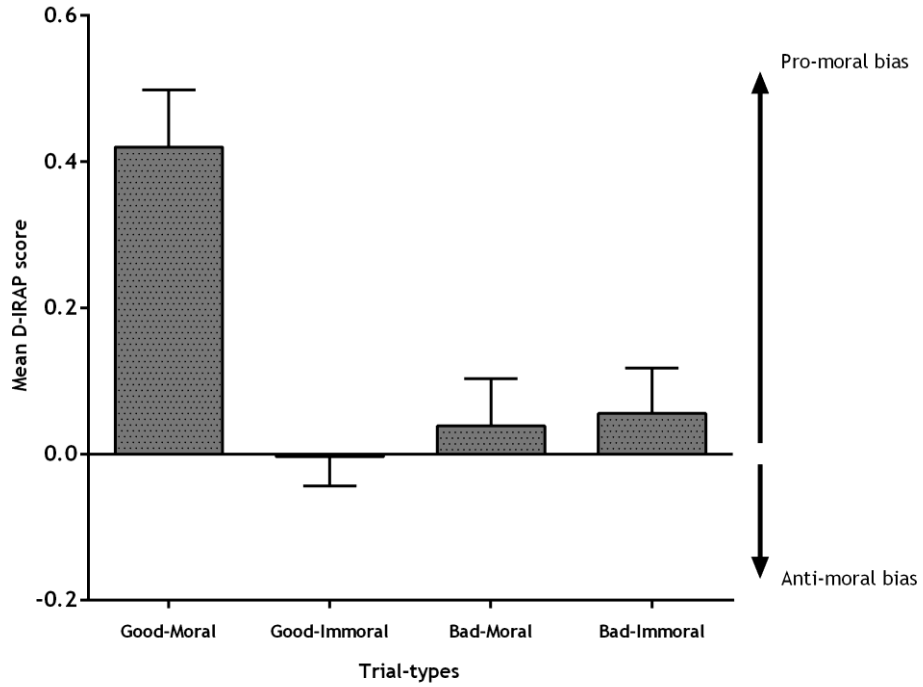


Figure 11. Mean D-IRAP scores with standard error bars for the four trial types of the CM-IRAP in experiment 3. Positive numbers indicate a pro-moral bias, and negative numbers indicate an anti-moral bias

In the new DF-IRAP (see Figure 12), participants tended to respond "True" more quickly than "False" on the *Do-good/Feel-good*, *Do-bad/Feel-Good* and *Do-bad/Feel-bad* trial types, and respond "False" more quickly than "True" in the *Do-Good/Feel-Bad* trial-type. Four one-sample t tests revealed two significant effects, for the *Do-good/Feel-good* ($t = 6.029$, $p < 0.01$) and *Do-bad/Feel-bad* trial-types ($t = 4.314$, $p < 0.05$).

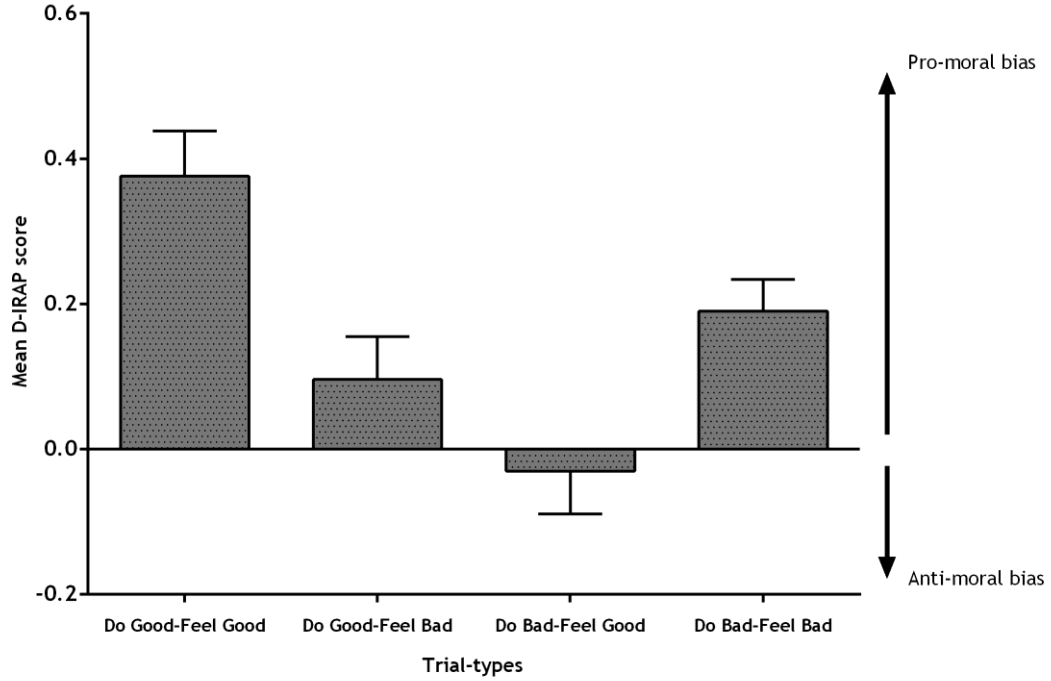


Figure 12. Mean *D*-IRAP scores with standard error bars for the four trial types of the DF IRAP in experiment 3. Scores above zero indicate pro-moral bias.

Predicting cheating behaviours. To determine whether the IRAPs could predict immoral behaviour in this study, Pearson correlations were calculated between the total MMT score and the mean *D*-IRAP scores for the four trial types in each IRAP (see Table 8, rightmost column). One statistically significant negative correlation emerged between the *Do-bad/Feel-bad* trial type and the total number of cheats in the MMT ($r = -0.37$, $p < 0.05$),

indicating that increasing pro-moral bias on the IRAP predicted less cheating. The correlation between the *Bad/Immoral* trial-type and the MMT that was significant in Study 1 approached significance in this study and showed the same trend, which is in line with our interpretation that people who engage in immoral behaviour might be more highly practised at denying it.

Correlations with psychopathy. We conducted Pearson correlations between the MMT score, the two indexes of psychopathy (primary and secondary) and the trial-types in both IRAPs. In the resulting correlation matrix (see Table 8) we observed a significant negative correlation between the DF-IRAP and the *Do-good/Feel-good* trial type and Secondary psychopathy (r = -0.38, $p < 0.05$), indicating that increasing levels of pro-moral bias on the IRAP predicted lower levels of self-reported situational psychopathy. Even though no statistically significant correlations emerged between the psychopathy subscales and the mean *D*-IRAP scores in the CM-IRAP, a moderate inverse correlation between the *Good/Moral* trial-type and Secondary psychopathy approached significance, indicating that increasing pro-moral bias predicted lower levels of self-reported situational and remorseful psychopathy.

Despite the goal being to correlate the IRAP trial-types and cheating and psychopathy, we also found that the MMT correlated with the secondary psychopathy factor (r = 0.38, $p < 0.05$), indicating that higher levels of situational psychopathy predicted increasing cheating responses.

Table 8.

Correlations between psychopathy, cheating and mean D-IRAP scores

	Primary Psychopathy		Secondary Psychopathy		MMT	
Trial-types	*r*	*p*	*r*	*p*	*r*	*p*
CM-IRAP						
Good/Moral	-0.113	> 0.1	-0.341†	0.07	-0.174	> 0.1
Good/Immoral	0.049	> 0.1	-0.174	> 0.1	0.076	> 0.1
Bad/Moral	0.113	> 0.1	-0.133	> 0.1	0.048	> 0.1
Bad/Immoral	0.277	> 0.1	-0.052	> 0.1	0.312†	0.09
DF-IRAP						
Do-Good/Feel-Good	-0.225	> 0.1	-0.384*	0.03	-0.083	> 0.1
Do-Good/Feel-Bad	-0.093	> 0.1	-0.176	> 0.1	0.104	> 0.1
Do-Bad/Feel-Good	-0.004	> 0.1	-0.251	> 0.1	-0.281	> 0.1
Do-Bad/Feel-Bad	0.092	> 0.1	-0.035	> 0.1	-0.374*	0.04

(*) Significant at the 0.05 level. (†) Approached significance.

Given that both the IRAPs and cheating correlated with secondary psychopathy we conducted partial correlations between trial-types and the two indexes of psychopathy with the MMT score as a controlling variable. Once again no significant correlations emerged between the trial-types and primary psychopathy, but secondary psychopathy was still associated with the *Good/Moral* trial-type ($r = -0.301$, $p < 0.04$) and the *Do-Good/Feel-Good* trial-types ($r = -0.382$, $p < 0.05$). This indicates that the IRAPs indeed seem to tap into verbal networks associated with boldness and impulsivity, conceptually associated with psychopathy, even when controlling for cheating.

Summary and Conclusions. The CM-IRAP effects in this study resemble those from previous studies. Again, *Good/Moral* and *Bad/Immoral* relations were consistent with common-sense expectations, although the size of the effect for the latter trial-type was relatively weak. No clear bias emerged from the *Good/Immoral* and *Bad/Moral* relations. The significant inverse correlation between the *Bad/Immoral* trial-type and the cheating measure was observed again (approaching significance). As before, this finding could be interpreted as indicating that increasingly strong claims that bad actions are immoral may reveal well established repertoires of deceptive behaviour.

The new deictic IRAP produced the expected, or common sense effects for the *Do-Good/Feel-Good* and *Do-Bad/Feel-Bad* trial-types. It also produced an inverse correlation between the *Do-Bad/Feel-Bad* trial-type and the cheating measure, indicating that participants who more strongly affirmed that they feel bad when engaging in bad behaviour tended to cheat less in the MMT. It is perhaps intriguing that the correlations between the MMT and the *Bad/Immoral* trial-type from the CM-IRAP and the *Do-Bad/Feel-Bad* trial-type from the DF-IRAP were in opposite directions. Specifically, this indicates that a bias towards confirming that bad actions are immoral predicts increased cheating, but a bias towards confirming that engaging in bad actions makes you feel bad predicts decreased cheating. Such a result seems highly counter-intuitive. One possible explanation could appeal to the basic assumption that the IRAP is sensitive to verbal histories. As noted previously, participants who engage in higher levels of cheating may well be better practiced at claiming that bad actions are immoral (as part of a general strategy to conceal immoral acts). Indeed, there are many social situations in which people may express strong opinions on how immoral or disgraceful or evil or awful a particular action might be. In contrast, the opportunities to talk about how you feel when you engage in an immoral action are far less frequent. In fact, most of us would likely avoid such situations – even thinking privately about how we feel when we behave badly is not something that many of us would embrace with any enthusiasm. Thus

verbal relations that involve condemning immoral behaviour may be quite distinct, functionally, from verbal relations associated with how one feels following an immoral act. We will return to this issue in the context of the General Discussion.

It is also worth noting that one trial-type from the CM-IRAP and one from the DF-IRAP produced correlations with secondary psychopathy that either were significant or approached significance, but no such evidence was obtained for primary psychopathy. Both correlations appear to be relatively intuitive in that they were all negative, indicating that lower levels of self-reported secondary psychopathy predicted increased pro-moral bias on the IRAPs. As mentioned earlier, secondary psychopathy refers to contextual, situational engagement in behaviours characterised by boldness and impulsiveness – however, secondary psychopathy does not feature the distinct lack of empathy and remorse that primary psychopaths present (Dean et al., 2013). It appears, therefore, the IRAPs employed in the current study were capturing behavioural repertoires that overlapped to some extent with so-called boldness and impulsiveness in the natural environment rather than lack of empathy and remorse. Again, we will also return to this issue in the General Discussion.

At this point in the current research programme, we had the possibility of addressing the issue of ecological validity, the lack of which is rather prevalent in many social sciences, including psychology (Henrich, Heine, & Noren, 2010). Most perspectives in moral psychology have been developed with data from European samples, whose moral judgments and choices are not necessarily representative of other populations, and thus their conclusions about morality have limited applicability outside those cultures. With the aim of providing some ecological validity to our results, we took advantage of the possibility of replicating Study 3 on a Colombian student sample from the main researcher's home university.

Study 4. Establishing Ecological Validity for Study 3

As Henrich et al. (2010) have noted, a large base of psychological research has been performed with American undergraduate students, whose views and behaviours are particular to their culture and might not extrapolate very well to other populations. Indeed, a number of studies have suggested that traits such as individualism, independence, and value attributed to choice and options, amongst others, are different in American samples than in other western samples (Henrich et al., 2010; Morling & Lamoreaux, 2013). Hence, an agenda for behavioural research with higher external validity demands cross-cultural studies or replications with samples from other cultures, in order to detect other factors where similarities or differences might be identified. With the dual aim of verifying whether our results up to this point showed consistency and reliability, and of contrasting them with results obtained from a different culture, we took advantage of the possibility of replicating Study 3 on a Colombian student sample from the main researcher's home university.

Materials and Methods

Participants. A convenience sample of 33 students from the Pontifical Xavier University in Bogotá, Colombia, participated in this study. Three participants did not meet criteria for the test blocks in either of the IRAPs and their data were consequently excluded from analysis. The final sample consisted therefore of 30 students (73.3% female) with ages between 18 and 23 (M = 18.93, SD = 1.25).

Measures. As this was a replication of Study 3, with the goal of providing convergent validity and exploring cultural differences, we used the same tasks. However, we conducted a cross-cultural validation procedure of the IRAPs and the consent forms, since we were working with a Spanish-speaking sample. We used the Spanish version of the Levenson Psychopathy Scale by Redondo (2012), which has a reported Cronbach's alpha of 0.77.

Cross-cultural validation procedures. Items in psychological instruments usually make reference to certain cultural backgrounds and scoring is based on behaviours prevalent in those backgrounds (Leong & Lyons, 2010). Plain translations might therefore fail to acknowledge cultural subtleties and yield inaccurate information because of the use of a culturally inappropriate interpretative framework. However, there is no standardised procedure for cultural adaptation of psychological instruments (Epstein, Santo, & Guillemin, 2014).

The main strategy to ensure the equivalence of items and procedures is to perform an initial translation into the target language, followed by a backward translation into the original (which should yield a version clearly equivalent to the original), and adjust the translation according to the findings (Callegaro Borsa, Figuereido Damásio, & Ruschel Bandeira, 2012). The translation phase is followed or done at the same time as the cultural adaptation phase, which involves adjusting items to culturally-specific practices. An example of this is an item in the Health Assessment Questionnaire that asked participants if they were able to sit in their bathtub. In Thailand, bathtubs are not common, so the Thai adaptation of this item asks about the ability to sit to pay homage to a sacred image (Epstein et al., 2014).

Despite the lack of standardisation, it seems that most methods achieve comparable results (Epstein et al., 2014). Our method to minimise the effect of linguistic and cultural biases on our results involved performing a process of translation-backtranslation-cultural assessment of the IRAP stimulus sets. The forward translation was performed by the main researcher, and the backtranslation was performed by another bilingual psychologist with experience with cross-cultural validation. A third psychologist then joined the group, and together they proceeded to examine the two versions and consensually solved the few conflicts that arose, in order to reach a final version. The three translators had Colombian

Spanish as their native language and good command of the English language and had lived in an English-speaking culture for more than one year.

IRAPs. The resulting stimulus sets from the cross-cultural validation procedure were as follows: in the CM-IRAP, "I think it is good to" was rendered as "*Creo que es bueno*" and "I think it is bad to" as "*Creo que es malo*". The positive labels were: "Ser moral", "Decir la verdad", "Cumplir la ley", "Ser justo", "Portarse bien", and "Ser honesto". The negative labels were "Ser inmoral", "Decir mentiras", "Romper la ley", "Hacer trampa", "Portarse mal", "Ser deshonesto". The rule for consistent blocks was "*Por favor contesta COMO SI las buenas acciones fueran BUENAS y las malas acciones fueran MALAS*", and the rule for the inconsistent blocks was "*Por favor contesta COMO SI las buenas acciones fueran MALAS y las buenas acciones fueran BUENAS*".

In the DF-IRAP, labels for the "*Do-Good*" trial-types were "Cuando me porto bien", "Cuando soy moral", "Cuando cumplo la ley", "Cuando sigo las reglas", "Cuando hago lo correcto", "Cuando digo la verdad". The label set for the "*Do-Bad*" trial-types consisted of the following phrases: "Cuando digo mentiras", "Cuando hago trampa", "Cuando quiebro la ley", "Cuando rompo las reglas", "Cuando soy inmoral", "Cuando engaño". The label that signalled a "*Feel-Good*" trial type was rendered as "me siento bien" and the "*Feel-Bad*" trial-type had the label "me siento mal". The rule for consistent blocks in this IRAP was "*Por favor contesta como si portarte BIEN te hiciera sentir BIEN y portarte MAL te hiciera sentirte MAL*" and the rule for inconsistent blocks was "*Por favor contesta como si portarte BIEN te hiciera sentir MAL y portarte MAL te hiciera sentirte BIEN*".

Procedure. Experimental sessions took place in a module at the laboratory of the Faculty of Psychology at the Pontifical Xavier University. Participants were greeted, thanked for volunteering and attending, and given similar instructions to those from previous studies – specifically that that they were about to take part in a preliminary study, designed to

evaluate different types of tests in order to select the most appropriate ones for future research. They were instructed to read a translated version of the Consent Form presented in Appendix E and reassured that their participation was voluntary, confidential and anonymous. After answering whatever questions participants had, they signed the consent form and were given COP$ 10.000 (roughly €5 at the time) as a token of appreciation. Then they sat in front of a Dell® OptiPlex™ 755 computer, running Windows XP and equipped with a standard 14 inch screen, and were presented with the MMT, with instructions delivered by the main researcher in Spanish (the task itself was not translated as it only required participants to respond to numbers). Upon finishing the task, they proceeded to complete the two IRAPs and the LPS. The experimental session ended with a full debriefing procedure. All participants reported understanding the reason for the deception involved in the MMT and agreed to have their data included in the analysis.

Results and Discussion

Cheating measure. Cheating levels in the modified MMT were similar to those observed in previous studies. Table 9 presents the percentage of participants who availed of the opportunity to cheat in each block, and the average number of cheating responses with its standard deviation and range. In general, the figures are only very slightly higher to those from Studies 1 and 3, and closer to what has been found in previous literature as discussed in Study 1.

Table 9

Descriptive statistics for the MMT.

Task	% who cheated	Average cheats	*SD*	Range
Slow block only	57.6	1.00	1.05	0-3
Fast block only	70.0	1.53	1.40	0-6
Combined	70.0	2.56	1.99	0-7

Psychopathy. Descriptive statistics for the Levenson Psychopathy scale are presented in Table 10. In general, the figures were similar, if only slightly higher, than in the previous study.

Table 10.

Descriptive statistics for the Levenson Psychopathy Scale.

Sub-scale	Mean	*SD*	Range
Primary Psychopathy	35.56	5.19	26-47
Secondary Psychopathy	25.40	4.91	18-36
Total score	60.96	7.95	46-79

IRAPs. The CM-IRAP showed similar trends to those from previous studies, in that participants seemed to respond "True" more quickly than "False" to the *Good/Moral* and *Bad/Immoral* trial types, but were quicker to respond "False" rather than "True" in the *Good/Immoral* and *Bad/Moral* trial-types. One-sample t tests revealed significantly different effects from zero for the *Good/Moral* ($t = 5.874$, $p < 0.01$) and *Bad/Immoral* ($t = 5.863$, $p < 0.05$) trial types. The effect for the *Bad/Moral* trial-type approached significance ($t = 1.823$, $p = 0.07$). The mean D-IRAP scores are presented in Figure 13.

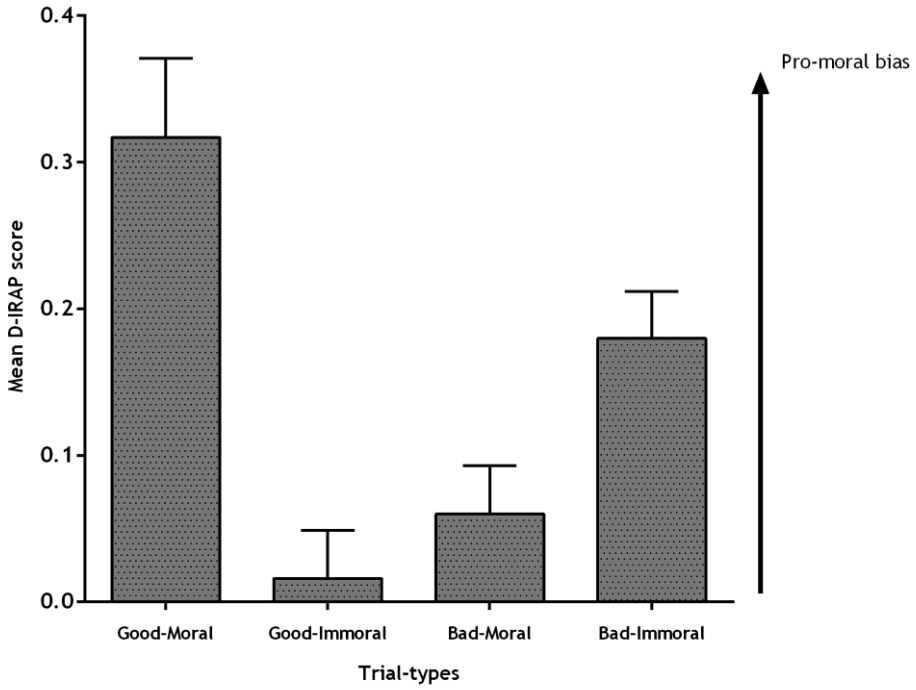


Figure 13. Mean *D*-IRAP scores with standard error bars for the four trial types of the FM IRAP in experiment 4. Scores above zero indicate pro-moral bias

The mean *D*-IRAP scores for the DF-IRAP are displayed in Figure 14 . Similar to the previous study, participants responded "True" faster than "False" on the *Do-Good/Feel-Good* and *Do-Bad/Feel-Bad* trial types. In this case, however, the response patterns for the remaining trial-types were reversed: whereas in Study 3 participants tended to respond "True" faster than "False" on the *Do-Bad/Feel-Good* trial-type and "False" faster than "True" on the *Do-Good/Feel-Bad* trial-type, the opposite was true here. One sample *t*-tests detected effects significantly different from zero for the *Do-Good/Feel-Good* ($t = 12.941$, $p < 0.01$), *Do-Bad/Feel-Good* ($t = 5.853$, $p < 0.01$) and *Do-Bad/Feel-Bad* ($t = 8.136$, $p < 0.01$) trial-types.

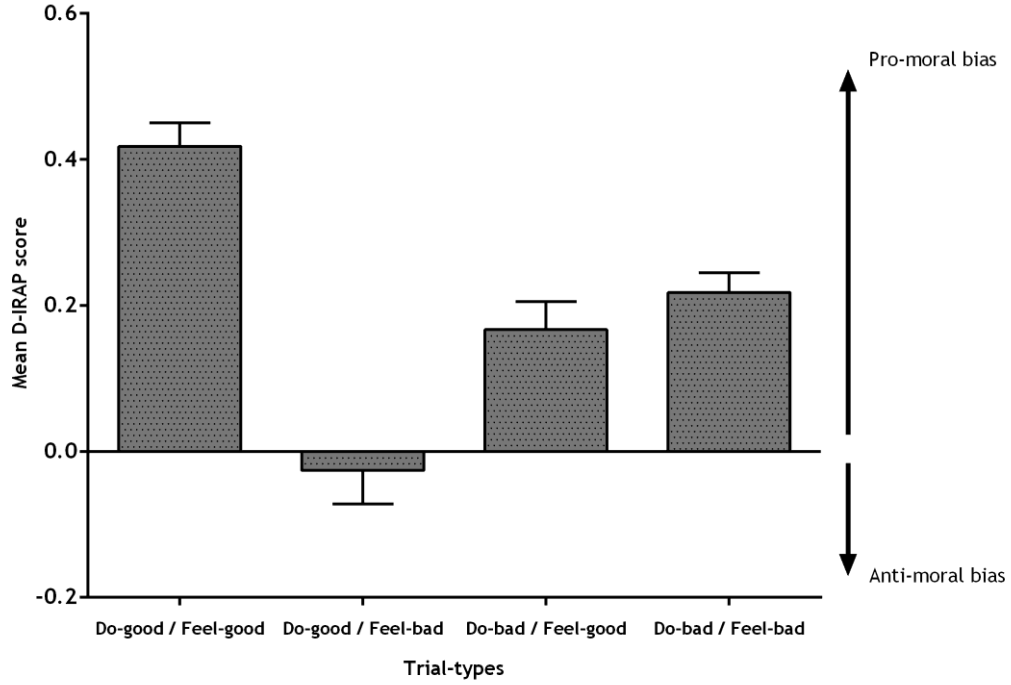


Figure 14. Mean *D*-IRAP scores with standard error bars for the four trial types of the FM IRAP in experiment 4. Scores above zero indicate pro-moral bias.

Predicting cheating behaviours. Pearson correlations between the mean D-IRAP scores in the trial types in each IRAP were entered into a correlation matrix with the psychopathy subscales and the cheating measure (see Table 11). In the CM-IRAP, a significant correlation between the *Bad/Immoral* trial-type and the MMT emerged, indicating that an increasing pro-moral bias in this trial-type predicted higher cheating. The same correlation was marginally significant in Study 3, but in the same direction. For the first time, a weak positive correlation between the *Good/Moral* trial-type and the MMT seemed to appear, although it did not reach significance. Our previous studies had very weak and negative, non-significant correlations between this trial type and the MMT.

In the DF-IRAP, the correlation between the *Do-Bad/Feel-Bad* trial-type and the MMT from Study 3 was observed (r = -0.373, $p < 0.05$), indicating that an increasing pro-moral bias predicted less cheating. Interestingly, the *Do-Bad/Feel-Good* trial-type showed a weak

positive correlation with the MMT, although it did not reach statistical significance – this trend indicates that participants who more strongly affirmed feeling good when engaging in bad actions also tended to cheat more. This latter result is unique to the current study.

Table 11.

Correlations between psychopathy, cheating and mean D-IRAP scores

	Primary Psychopathy		Secondary Psychopathy		MMT	
Trial-types	*r*	*p*	*r*	*p*	*r*	*p*
CM-IRAP						
Good/Moral	-0.245	> 0.1	-0.040	> 0.1	0.308†	0.09
Good/Immoral	-0.120	> 0.1	0.122	> 0.1	0.037	> 0.1
Bad/Moral	-0.202	> 0.1	-0.259	> 0.1	-0.027	> 0.1
Bad/Immoral	-0.212	> 0.1	-0.013	> 0.1	0.389*	0.03
DF-IRAP						
Do-Good/Feel-Good	-0.259	> 0.1	-0.339†	0.06	-0.036	> 0.1
Do-Good/Feel-Bad	0.268	> 0.1	0.090	> 0.1	-0.141	> 0.1
Do-Bad/Feel-Good	0.086	> 0.1	-0.156	> 0.1	0.331†	0.07
Do-Bad/Feel-Bad	-0.248	> 0.1	-0.114	> 0.1	-0.373*	0.04

(*) Significant at the 0.05 level. (†) Approached significance.

Correlations with psychopathy. A marginally significant negative correlation emerged between the *Do-Good/Feel-Good* trial-type in the DF-IRAP and secondary psychopathy, which indicates that an anti-moral bias predicts higher situational and emotive psychopathy. A similar correlation was found in Study 3, although it failed to reach significance here at the $p < 0.05$ level. Also similar to Study 3, the secondary psychopathy subscale correlated with the

MMT (r = -0.358, p = 0.05), indicating that higher levels of situational psychopathy predicted increasing cheating responses. All other correlations were non-significant.

Summary and conclusions. This replication showed similar effects to the previous study. The distribution of mean *D*-scores in both IRAPs points to faster and more accurate responding that confirms common sense expectations, especially on the *Good/Moral* and *Bad/Immoral* trial-types in the CM-IRAP and on the *Do-Good/Feel-Good* and *Do-Bad/Feel-Bad* trial-types in the DF-IRAP.

However, and in line with findings from the first three studies, the IRAPs also identified less clear biases in certain trial-types, perhaps reflecting their sensitivity to less well-established verbal relations. This was especially true for the *Good/Immoral* trial-type in the CM-IRAP, for which we failed to find significant effects throughout the studies up to this point. It appears, therefore, that participants had difficulties denying (more readily than confirming) that good actions are immoral. In contrast, the *Bad/Moral* trial-type showed a significant effect in Study 1 and an effect that approached significance in the current study, indicating a readiness to deny (more readily than confirm) that bad actions are moral. We will come back to this point and suggest possible explanations for this pattern of results in the context of the General Discussion.

Similar to Study 3, the DF-IRAP produced the expected and relatively strong effects for the *Do-Good/Feel-Good* and *Do-Bad/Feel-Bad* trial-types, and this last trial-type was able to predict cheating in the MMT consistently across the two studies presented in the current chapter. The *Do-Good/Feel-Bad* trial-type failed to yield significant effects in both studies but the *Do-Bad/Feel-Good* trial-type produced a relatively strong and significant effect in the current study (but not in the former experiment). At the present time, it remains unclear why we obtained this difference across the two studies. Of course, it could be due to simple error

variance or some undefined cultural differences between Irish and Colombian participants. Again, we shall return to this issue in the context of the General Discussion.

At this point in the current research programme, the CM-IRAP has shown its ability to capture the expected, common-sense effects of morality of good actions and immorality of bad actions. It was also able to predict cheating to a certain degree through its *Bad/Immoral* trial-type, leading to our interpretation that stronger confirmations that bad actions are immoral might sometimes indicate a higher likelihood of engaging in immoral actions. Nevertheless, we have found the DF-IRAP to be a more reliable predictor, both in conceptual and empirical terms, in that the correlations were relatively strong and intuitively predictable (i.e., confirming Do-Bad/Feel-Bad relations more quickly than denying this relation predicted lower levels of cheating). At this point, therefore, it appears that the IRAP that targeted deictic relations (i.e., how the participant feels about their own moral and immoral actions) was a more appropriate tool for assessing implicit morality than an IRAP that targeted implicit attitudes about morality *per se.* Therefore, for the last part of the current thesis, we decided to use the DF-IRAP as the single implicit measure of deceptive behaviour.

Chapter 6.

Intervening to curb deceptive behaviour

Intervening to curb deceptive behaviour

Unethical behaviour has direct significant institutional and personal effects. The financial effects of immorality have been widely studied thanks to the severity of their consequences and their readily measurable impact. For example, the collapse of Enron Corporation in 2004, brought about by shady business practices and lax audit processes, resulted in the loss of thousands of jobs and the evaporation of shareholder value and retirement plans. Gino, Schweitzer, Mead and Ariely (2011) point out that around one trillion dollars is lost in the economy of the United States through diverse forms of immoral behaviour.

There are also psychological consequences to immorality. Acting immorally seems to evoke feelings of guilt, produce discomfort upon recall or mention of past unethical actions and make moral agents believe the lies they tell (Tobey Klass, 1978). In a study with early career lawyers, Kammeyer-Mueller, Simon and Rich (2012) found higher levels of emotional exhaustion and decreased career satisfaction in participants who felt pressed by their employer to engage in practices that countered their own sense of morality. And deception, even if undiscovered, has been found to generate distrust and wreak havoc in romantic relationships (Sagarin, Rhoads, & Cialdini, 1998).

Aware of the pervasiveness of immoral behaviour and its costs, governments, companies and training facilities worldwide routinely offer workshops and courses intended to decrease cheating, deception and harmful unethical practices. However, these interventions tend to be based upon traditional conceptions of morality as an inner sense responsible for behaviour that develops in a predictable sequence towards an ultimate goal. As such, they target vague internal constructs such as “moral self-concept” or “moral reasoning processes”. However, this focus on inferred constructs normally puts them at high risk of failing to identify and address functional determinants of immoral behaviour.

Zhang et al. (2014) propose that interventions to reduce unethical behaviour can be structure-oriented or values-oriented. Structure-oriented interventions aim to reduce external temptations to cheat and increase rewards for ethical behaviour. This includes the implementation of policies, increasing the likelihood and the size of punishment for immorality, increasing rewards for ethical behaviour, and generally designing environments in such a way that unethical behaviours are impossible or unlikely. For example, strategic placement of CCTV cameras and access control barriers reduced car theft by up to 85%, sometimes even eliminating it completely, in several cities in the UK and the USA (La Vigne & Lowry, 2011).

Values-oriented interventions, on the other hand, target a personal desire to behave ethically. For instance, reminding people of their own immoral actions in the past has been suggested to produce compensatory moral action in the form of increased prosocial behaviour and less cheating (Jordan et al., 2011). Other possibilities are to expose people to general positive values, which has been found to correlate with more prosocial attitudes and actions (Zhang et al., 2014) and to promote inferences of one's self-concept to be more ethical (Aquino & Reed, 2002). Finally, priming tasks have found that exposing people to positive concepts might help curb immoral behaviour. For example, Chugh, Kern, Zhu and Lee (2014) asked participants to remember situations in which they felt secure and accepted, or anxious and rejected, and found decreased deception in the first (secure) group. Priming with concepts that had a less clear association with cheating (for example "time" instead of "money") seems to produce less unethical behaviour as well (Gino & Mogilner, 2014).

Within the context of experimental social psychology, many researchers have sought out interventions that might help reduce the effects of stereotyping, prejudice, racism, and other socio-cognitive processes. A large set of studies (Greenberg, Pyszczynski, Solomon, Simon, & Breus, 1994) support the notion that thinking of one's death temporarily changes

people's perceptions and reactions towards threats or challenges to personal worldviews. These "mortality salience" effects have been shown to permeate diverse social processes such as interpersonal attraction, obedience, nationalism, and it has been proposed that their mechanism of action is the temporary perception of a threat to a stable worldview, responsible for an increase in negative affect. Based on these reported mortality salience effects, we decided to test two premises: that bringing attention to one's mortality causes cheating to decrease, and that the DF-IRAP will be sensitive to those changes and will continue to be useful as a tool for implicit evaluation of morality-related verbal networks. Testing these hypotheses was the main goal of the two studies presented in the current chapter.

Study 5: The Effect of a Values-Oriented Intervention on Cheating Behaviour

Death is the unavoidable, ultimate fate of all human beings. Intuitively, such experience should be regarded as a necessary component of life and embraced as such, but both everyday experience and scientific research (pioneered by Becker, 1973) suggest that reacting to thoughts of death with anxiety, fear and other negative emotions is commonplace in many cultures. To account for these reactions, Greenberg et al. (1994) developed Terror Management Theory (TMT), which proposes that awareness of mortality encourages a desire for finding meaning and building self-esteem as a form of protection against the threat of eventual death. The methodological device used to test this idea in the laboratory is the Mortality Salience Induction (MSI), a task that encourages participants to think and reflect on their own death, and thus bring it to consciousness and make it more salient.

The MSI requires participants to respond to four open-ended statements related to thoughts of their own death (the items are presented in Appendix H), and it is normally presented as part of a package of tasks. The induction has consistently shown a set of so-called "mortality salience effects" that play a role in many different social situations and issues: in a meta-analysis of more than 160 studies that used the MSI, Burke et al. (2010) found significant effects of the task on such diverse measures as state guilt, desire for control, attitudes towards animals, evaluations of others, moral relativity, preference for positive words, willingness to interact, and many others.

The proposed mechanism for the mortality salience effects is the temporary perception of a threat to a stable worldview, responsible for an increase in negative affect. The task has been found to heighten anxiety and induce negative emotions (Routledge et al., 2010), but has also been shown to have a wealth of positive effects, such as an increase in reciprocity (Schindler, Reinhard, & Stahlberg, 2013), prosocial orientation (Niesta, Fritsche, & Jonas, 2008), intentions to engage in healthy behaviours (Arndt, Schimel, & Goldenberg,

2003; Bevan, Maxfield, & Bultmann, 2014), and likelihood to participate in donation appeals (F. Cai & Wyer, 2014). In any case, the duration of the effect seems to be minutes or hours (Burke et al., 2010).

Therefore, in this study, we sought to determine whether the MSI had an impact on engagement in cheating behaviour and whether the DF-IRAP is sensitive to this impact.

Materials and Methods

Participants. Fifty-five students (64% females) recruited from the Departmental Volunteer pool from the National University of Ireland, Maynooth, participated in this study, with ages between 18 and 26 (M = 19.74, *SD* = 2.21). They were recruited into an intervention and a control group (details of the tasks will be provided below). Analyses proceeded with data from 50 participants because five did not achieve test criteria in the practice blocks of the IRAP, and their data were consequently excluded from the dataset.

Measures. We used a variation of the Mortality Salience Induction (Greenberg et al., 1994). The general goal of the Mortality Salience Induction and its numerous variations is to get participants to reflect on their own mortality. The original intervention simply asks participants four open-ended questions about their own mortality, but in the version used in this study, we also used a priming task that helped strengthen reflections on mortality in other studies conducted in our laboratory (Hussey & Barnes-Holmes, 2015). This latter task involves presenting a piece of paper containing a number of dots that matches the expected number of weeks left to live for participants, according to their age and gender, and making them aware of the fact that it is a relatively low number of dots. The complete script and questions for the intervention and control versions are presented in Appendix H. Based on the results of Studies 3-5, we selected the DF-IRAP to assess the effect of the intervention.

Procedure. Sessions took place in experimental cubicles identical to the ones used in previous studies. Participants were given similar information to that provided in other experiments in the current thesis, i.e., that this was a preliminary study designed to evaluate different types of tests in order to select the most appropriate ones for future studies. This was followed by them reading and signing the informed consent form (Appendix E), and reassurance by the researcher that their participation was confidential, voluntary, anonymous and that they could withdraw from the study at any time, which none of them decided to do. They were then given €5 for their participation that they could keep even if they decided to terminate their participation mid-way. Before starting the data collection proper, participants were asked if they had experienced bereavement over that past 12 months or if they had a history of diagnosed psychiatric disorders. No volunteers reported either, so the data for the full sample was included in the analyses, except for the five mentioned participants who failed to reach the test blocks on the DF-IRAP. The experimental session itself started with either the Mortality Salience Induction (MSI) task or the Control task.

Mortality Salience Induction (MSI) and debriefing. If the participant had been assigned to the intervention group, the researcher verbally delivered the following information:

> "I think it's often very easy to forget just how short life is, especially for young healthy students. To help convey this, I've put together this diagram for you. Given that I know your age and gender, it's trivial for me to estimate your expected lifespan"

At this point, the researcher placed in front of the participant a sheet of paper containing small dots arranged in a square in the centre of the sheet, measuring approximately 2.2 inches and containing 52 rows and 52 columns, for a total of 2.704 dots. The researcher then explained:

> "Based on that, the number of dots on this piece of paper is equal to the number of weeks you have left to live... [long pause] I promise that I'm not trying to trick you or deceive you – it is a surprisingly small number of dots, isn't it? The thing about dots is that once you spend them, you cannot get them back. This is not a rehearsal, you will not get a second shot. This is your life, right now, ending, one day at a time. [pause] The other thing about dots is that they run out, no matter what you do. Make no mistake, death *is* coming. You have a limited number of days left on this planet, and like all of us, you're faced with the difficult question of what you're going to do with them. [pause] How many of these dots will be well spent dots, doing things that you truly value, like time with friends and family, and how many dots will be more like hovering dots and X-factor dots?"

A short pause followed, after which the researcher gave the following instruction "With all that in mind, I'd like you to write out a few lines about what you think dying itself will be like", and put another sheet of paper in front of the participant, that contained the following four items:

1. "What emotions does the thought of your own death arouse in you?"
2. "Jot down, as specifically as you can, what you think will happen to you physically as you die and once you are physically dead"
3. "The one thing I fear most about my death is..."
4. "My scariest thoughts about my death are..."

After ensuring that the participant had understood the instructions, the researcher left the room and the participant started the task. Upon its completion, the researcher reentered the room and told that some people found the task to be unpleasant, given that death is not something people generally like to discuss or bring to consciousness. However,

they were told, some research suggested that reflecting on one's death actually brought some positive effects, like increased pro-social disposition and motivation, probably due to the acknowledgement that life eventually ends and that we have limited time to do what we value. In this short conversation, the researcher also aimed to determine whether the participant was experiencing anxiety or stress levels beyond what could be reasonably expected. Many participants, however, reflected positively on the experience and acknowledged that it seems likely that awareness of mortality helps put things in perspective and increase the chances of engaging in valued action.

A protocol designed to ensure the well-being of the volunteers was in place, that involved terminating the experiment immediately, accompanying the participant to the University Medical Centre, reporting the incident to a member of staff at the Department of Psychology, and follow up (National University of Ireland, Maynooth, 2015). However, no participant reported having experienced heightened psychological distress and therefore the protocol did not need to be activated.

Control task. The procedure for participants assigned to the control group was similar, but the control task aimed to replace mortality-related stimuli with neutral words and expressions that had nothing to do with values. The script was as follows, with departures from the original underlined:

> I think it's often very easy to forget just how <u>big our solar system is</u>, especially for young students. To help convey this, I've put together this diagram for you. [place sheet in front of them]. The number of dots on this piece of paper <u>is equal to the number of million kilometres that separates the Sun and the Earth</u>. [said very slowly and carefully, and then a long pause]. I promise that I'm not trying to trick or deceive you. It's a surprising number of dots, isn't it? I'd like you to write out a few lines about what you think <u>about the size of our solar system</u>.

1. What emotions does the thought of the size of our solar system arouse in you?
2. Jot down, as specifically as you can, whether you think mankind will be able to travel throughout our solar system.
3. "The one thing that comes to mind when pondering the size of our solar system is…"
4. "My thoughts about our the size of solar system are…"

MMT and IRAP. The first task was immediately followed by the MMT and the DF-IRAP, delivered in the exact same manner as in Study 3. Afterwards, participants were thoroughly debriefed according to the procedure that has been used in the previous studies. All participants agreed to have their data included in the analysis.

Results and Discussion

Cheating measure. Descriptives for the cheating measure (see Table 12) showed that overall 58% of the participants availed of the opportunity to cheat – a slight decrease from the numbers from previous studies, and the lowest figure from the set of studies using the MMT up to this point. The average number of cheats per group was lower for the MSI in every case, supporting our hypothesis, although independent t-tests failed to find significant differences between the two groups in every case ($p > 0.1$).

Table 12.

Descriptive statistics for the MMT by task in Study 5

	MSI		Control	
MMT Block	% cheaters*	Avg. cheats (SD)	% cheaters*	Avg. cheats (SD)
Slow	20.0	0.200 (0.408)	32.0	0.360 (0.569)
Fast	56.0	1.080 (1.382)	52.0	1.480 (1.782)
Combined	56.0	1.240 (1.562)	56.0	1.840 (2.035)

(*) Percentage reported within task

IRAP. Figure 15 presents the mean *D*-IRAP scores for the DF-IRAP per group. In general, participants responded "True" more quickly than "False" on the *Do-Good/Feel-Good* and *Do-Bad/Feel-Bad* trial-types, and the opposite responding pattern was observed in the *Do-Good/Feel-Bad* and *Do-Bad/Feel-Good* trial-types. One-sample t-tests showed effects significantly different from zero at the $p < 0.02$ level for every trial type. A 2x4 repeated measures ANOVA with the trial-types as within-subjects variables and the task as a between-subjects factor confirmed a statistically significant effect of the task on the mean *D*-IRAP scores ($F = 4.092$, $p < 0.05$). In order to determine the direction and location of the effect, independent t-tests grouped by Task (MSI or Control) were conducted. These revealed that the MSI group scored higher ($M = 0.480$, $SD = 0.292$) than the Control group ($M = 0.304$, $SD = 0.282$) in the *Do good-Feel-good* trial type ($t = 2.160$, $p < 0.03$, $g_S = 0.602$ [95% *CI*: 0.03 - 1.17]). The effect for the *Do-Good/Feel-Bad* trial-type suggested by the figure was marginally significant ($t = 1.693$, $p = 0.09$). No significant effects were found for the remaining trial-types.

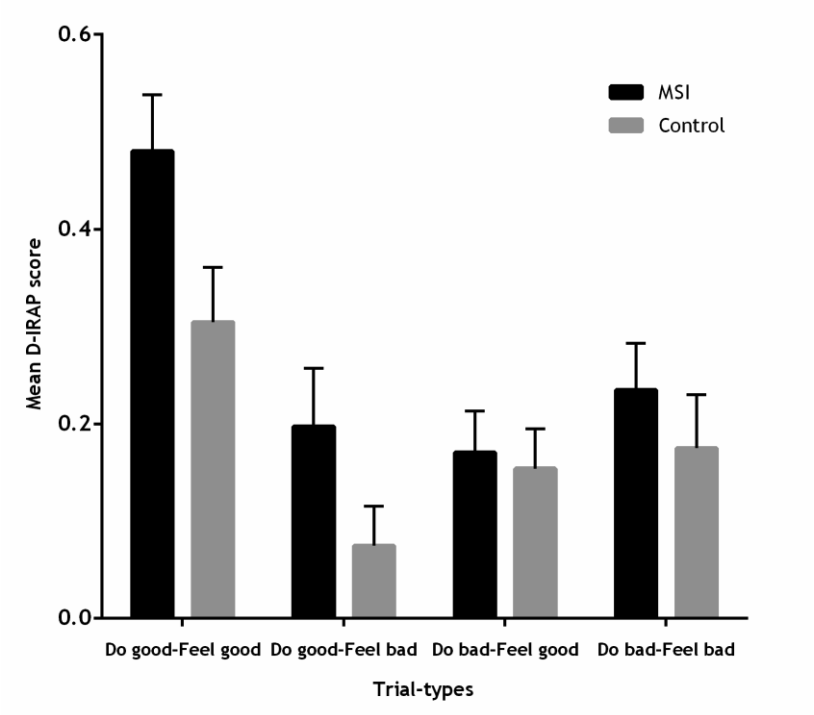


Figure 15. Mean *D*-IRAP scores with standard error bars for the four trial types of the DF-IRAP, divided by group (MSI or control).

Predicting cheating. As observed in Table 13, which shows the correlation matrix between trial-types in the DF-IRAP and the MMT score, divided by group (MSI or Control), the *Do-Bad/Feel-Bad* trial-type seemed to maintain its predictive power on the final cheating score, slightly improving on the results from previous studies. The overall correlation between this trial-type and the MMT (r = -0.414, $p < 0.01$) suggests once again that this trial-type is probably the best predictor of cheating scores. The *Do-Good/Feel-Bad* trial-type showed a weak to moderate negative correlation with the MMT that approached significance, indicating that a pro-moral bias in this trial-type predicts lower cheating.

Table 13

Correlations between the trial-types in the DF-IRAP and cheating per group

	MSI		Control	
	r	*p*	*r*	*p*
Do-Good/Feel-Good	-0.311	< 0.1	-0.208	> 0.2
Do-Good/Feel-Bad	-0.345†	0.09	-0.319	> 0.1
Do-Bad/Feel-Good	-0.113	< 0.5	-0.019	> 0.5
Do-Bad/Feel-Bad	-0.398*	> 0.05	-0.406*	< 0.05

Summary and conclusions. The distribution of the DF-IRAP in this study was similar to what has been observed in the previous experiments, in that participants readily confirmed that good actions make them feel good and bad actions make them feel bad - all the trial-types are in fact in the expected direction. Group comparisons between the MSI and control task showed that the MSI seemed to have a strong effect on the *Do-Good/Feel-Good* trial-type, indicating that participants who completed the mortality salience intervention confirmed more quickly and accurately that engaging in good actions made them feel good. In

fact, the trend was that the MSI produced slightly higher effects than the control task in every trial-type, although the difference was statistically significant only for the aforementioned *Do-Good/Feel-Good* trial-type. It is worth noting that a difference between the *Do-Good/Feel-Bad* trial-type approached significance, which indicates that participants who completed the MSI denied more strongly that engaging in good actions made them feel bad.

We observed somewhat counterintuitive results when examining the results from the cheating measure. On the one hand, the MSI seems to have decreased the amount of cheating, if not by a large margin. We had hypothesised that this would be the case for the MSI, but a similar decrease was observed in the group presented with the neutral intervention. At this time, the reason for this general decrease remains unexplained but possible explanations will be addressed in the context of the General Discussion. Correlations between the cheating measure and the trial-types in the IRAP yielded significant inverse correlations with the *Do-Bad/Feel-Bad* trial-type, similar to what was observed in the studies from the previous chapter. A correlation that approached significance was found between the *Do-Good/Feel-Bad* trial-type and the MMT, indicating that stronger denials that engaging in good actions evoked bad feelings seemed to predict lower cheating.

At this point in the current study, the observed effect prompted us to perform a conceptual replication in order to determine whether the effect is stable, and to pinpoint the part of the task responsible for the effect. Therefore, in Study 6, we aimed to separate the two components of the intervention and determine whether they still produced the observed effect.

Study 6. Isolating the Components of the Previous Intervention

The MSI variation we used in Study 5 can be thought of as a two-component task, one being the dots part and the other being the questions. Even though both aim to increase the salience of mortality, the dots task also targets the idea of time heavily: participants are instructed to look at the dots and realise that they represent the time they have left to live.

In their study of priming and immorality, Gino and Mogilner (2014) found that priming people with time-related words, as opposed to money-related words, seems to decrease unethical behaviour in a deception task. They speculate that the idea of money is more associated with immorality in daily life than time. This led to the idea of testing the two components independently in order to determine what component of the MSI task we used in Study 5 carries the largest effect in the observed reduction in immorality.

Materials and Methods

Participants. A convenience sample of 56 Psychology students from the Pontifical Xavier University in Bogotá, Colombia was used for this study. They were quasi-randomly assigned to two intervention groups, the "dots" group and the "questions" group. The data for five participants had to be excluded from the analysis because they failed to reach the criteria for the test blocks in the IRAP, leaving a total of 51 participants, 26 in the dots group and 25 in the control group. 67.7% of the sample were females.

Measures. The MSI from Study 5 was separated into its two hypothesised components. Participants assigned to the "dots" condition were presented with the first part of the MSI (Appendix G, from "Dots section begins" until "Questions section begins"), and participants assigned to the "questions" condition were presented with the first part of the

Validation procedure for the MSI. As described when presenting Study 4, ensuring procedural equivalence is not an exact science and different approaches seem to yield similar results. Our choice in the matter has followed the work of other researchers in Latin America and involved a process of translation-backtranslation-cultural assessment of the scripts for the dots and questions tasks. Like before, the main researcher performed the forward translation into Spanish, and the same team responsible for backtranslation and cross-validation in Study 4 helped with the procedure in the same way. The questions themselves have been translated into Spanish by Campos Vizcarra (2013) and we have used them here. The complete scripts are presented in Appendix I.

Procedure. Data collection took place at the Psychology Laboratory at the Pontifical Xavier University, inside standard, insonorised experimental modules. Participants were welcomed to the laboratory, thanked for their willingness to participate, and given similar instructions to those from previous studies – specifically that that they were to evaluate different types of tests in order to select certain tasks for future research. They were presented with a translated version of the Consent Form from Appendix E and informed that their participation was voluntary, confidential and anonymous. After reading and signing the consent form, they were given COP$ 10.000 (roughly €5 at the time) as a token of appreciation.

"Dots" task. If the participant had been assigned to the Dots sub-task, the researcher verbally delivered the information presented in Appendix I. The task was introduced by delivering the following information (in Spanish):

> "Creo que suele ser muy fácil olvidar lo corta que es la vida, especialmente para estudiantes jóvenes y sanos. Para ayudarte a comprender esto, he creado este diagrama. Dado que conozco tu edad y tu sexo, puedo estimar fácilmente tu expectativa de vida" [*I think it's often very easy to forget just how short life*

is, especially for young healthy students. To help convey this, I've put together this diagram for you. Given that I know your age and gender, it's trivial for me to estimate your expected lifespan].

At this point the participant was given the sheet of paper containing dots, and told:

"Con base en eso, el número de puntos en esta hoja de papel representa el número de semanas que te quedan de vida [decirlo lenta y cuidadosamente, y luego una pausa larga). Te aseguro que no estoy tratando de engañarte con esto. Es un número sorprendentemente pequeño de puntos, ¿no es verdad? La cosa con estos puntos es que una vez que los gastas no los puedes tener de nuevo. Esto no es un ensayo, no habrá una segunda oportunidad. Esta es tu vida, ahora mismo, acabándose día a día. Otra cosa de estos puntos es que van a acabar, sin importar lo que hagas. La muerte llegará – no pienses que no. ¿Cuántos de estos puntos vas a gastar bien, haciendo cosas que de verdad valoras, como pasar tiempo con tu familia, y cuántos serán puntos haciendo pereza y viendo 'Yo me llamo'?" [*Based on that, the number of dots on this piece of paper is equal to the number of weeks you have left to live... [long pause] I promise that I'm not trying to trick you or deceive you – it is a surprisingly small number of dots, isn't it? The thing about dots is that once you spend them, you cannot get them back. This is not a rehearsal, you will not get a second shot. This is your life, right now, ending, one day at a time. [pause] The other thing about dots is that they run out, no matter what you do. Make no mistake, death is coming. You have a limited number of days left on this planet, and like all of us, you're faced with the difficult question of what you're going to do with them. [pause] How many of these dots will be well spent dots, doing*

things that you truly value, like time with friends and family, and how many dots will be more like hovering dots and X-factor dots?]

Participants were then instructed to take a few minutes to reflect mentally on the information provided, during which the experimenter left the room. After about a minute had passed, the experimenter reentered the module and started the debriefing described in Study 4.

Questions task. If the participant was assigned to this group, the sub-task was introduced in a similar way to the Dots task, without the critical manipulation, by delivering this message:

> "Creo que suele ser muy fácil olvidar lo corta que es la vida, especialmente para estudiantes jóvenes y sanos. Ahora te voy a pedir que reflexiones un poco sobre tu propia vida y su final inevitable, y que escribas en esta hoja las respuestas a las preguntas que están escritas" [*I think it's often very easy to forget just how short life is, especially for young healthy students. Now I will ask you to reflect a little bit on your life and its inevitable end, and that you write some answers for the questions on this sheet*].

This was followed by a short pause, after which the researcher prompted the participant to write out a few lines about their perceptions of what their death would be like, on a piece of paper with the following questions, equivalent to their English versions, on it:

1. ¿Qué emociones te genera pensar sobre tu propia muerte?
2. Escribe, con tanto detalle como puedas, qué crees que te pasará físicamente cuando mueras y cuando estés físicamente muerto(a)
3. "Lo que más me asusta de mi muerte es…"
4. "Mis pensamientos más aterrorizantes sobre la muerte son…"

Having made sure that the participant had understood the instructions, the researcher left the room and the participant started the assigned task. Independently of the task, the researcher reentered the room when the participant had finished and started the debriefing by telling participants that the task was unpleasant to some, because death is not a popular topic and certainly not something people think much about. They were told about the positive effects of the task and during this conversation the researcher aimed to ensure that the participant was not distressed or experiencing any negative affect. The session continued with the administration of the MMT and the IRAP in the same manner as in Study 4, and finished with the same debriefing procedure. All participants agreed to have their data included in the following analysis.

Results and Discussion

Cheating measure. The descriptives for the MMT are presented in Table 14. In general, the dots group had slightly lower levels of cheating, which is in agreement with results from Study 5 and our hypothesis that the intervention decreases immoral behaviour. However, the Questions group presents levels of cheating that are rather similar, if only slightly lower, to those found in previous studies, with more than half of the sample having cheated at least once. Therefore, from these results, the Dots task seemed to be the main carrier of the decrease in cheating. In every case, the average number of cheats on the MMT is lower on the Dots group compared to the Questions group, although the difference is not significant on an independent samples *t*-test ($p > 0.1$).

Table 14

Descriptive statistics for the MMT by task in Study 6

	Dots		Questions	
MMT Block	% cheaters*	Avg. cheats (SD)	% cheaters*	Avg. cheats (SD)
Slow	19.2	0.269 (0.604)	28.0	0.360 (0.700)
Fast	50.0	1.154 (1.592)	60.0	1.640 (1.705)
Combined	50.0	1.423 (1.943)	64.0	2.120 (2.186)

IRAP. The mean *D*-IRAP scores for Study 6, divided by group, are presented in Figure 16. The general distribution is strikingly similar to that from every study using the DF-IRAP so far. Participants confirmed faster and more accurately that doing good things evoked good feelings and that engaging in bad actions conversely produced bad feelings – the common sense expectation that the DF-IRAP has consistently shown so far. The remaining trial-types are also in the expected, pro-moral direction, and one-sample t tests to determine if the effects were significantly different from zero yielded significant results for all but the *Do-Good/Feel-Bad* trial-type ($t = 1.945$, $p = 0.06$ for the Dots group and $t = 1.503$, $p < 0.1$ for the Questions group). In order to determine whether the task had an effect on the mean *D*-IRAP scores, we conducted a 2x4 repeated measures ANOVA that confirmed a difference favouring the Dots task ($F = 4.054$, $p < 0.05$). This was followed by independent T tests that revealed that the Dots group scored higher ($M = 0.310$, $SD = 0.161$) than the Questions group ($M = 0.0.226$, $SD = 0.124$) on the *Do Good/Feel-Good* trial type ($t = 2.081$, $p < 0.05$, $g_S = 0.57$ [95% *CI*: 0.01 – 1.14]) and on the *Do-Bad/Feel-Bad* trial type ($t = 2.032$, $p = 0.05$, $g_S = 0.56$ [95% *CI*: 0.00 – 1.13]). No significant effects were found for the remaining trial-types.

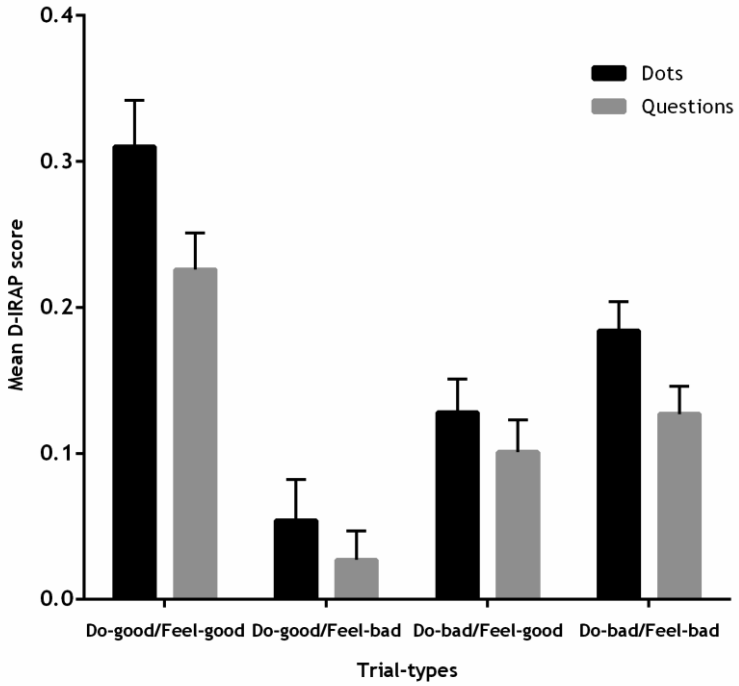


Figure 16. Mean *D*-IRAP scores with standard error bars for the four trial types of the DF-IRAP in Study 6, divided by group (MSI or control).

Predicting cheating. As depicted in Table 15, a pro-moral bias on the *Do-Bad/Feel-Bad* trial-type predicted lower cheating (r = -0.410, p = 0.03) in the Dots task only. None of the remaining trial-types predicted MMT scores, although the overall correlation between the *Do-Bad/Feel-Bad* trial-type and the cheating measure persisted when considering the entire sample (r = -0.389, p < 0.01).

Table 15

Correlations between the trial-types in the DF-IRAP and cheating per group

	MMT total score			
	Dots group		Questions group	
	r	*p*	*r*	*P*
Do-Good/Feel-Good	-0.324	< 0.1	0.215	< 0.1
Do-Good/Feel-Bad	-0.315	< 0.1	-0.273	< 0.1
Do-Bad/Feel-Good	-0.069	< 0.1	0.296	< 0.1
Do-Bad/Feel-Bad	-0.410*	0.03	-0.316	< 0.1

Summary and conclusions. The distribution of mean *D*-IRAP scores in the DF-IRAP in the current study was similar to that from previous studies, which is further evidence that the IRAP is a useful tool that can capture relational responding related to morality. The trial-types were all in the expected direction, although the *Do-Good/Feel-Bad* trial-type once again exhibited some inconsistency, which supports our notion that opportunities to exercise this particular type of response throughout common verbal histories are limited, and this is reflected on the performance on the IRAP.

Our deconstruction of the MSI task that we had used in Study 5 suggested that the Dots task, which makes participants reflect on their mortality by strongly targeting the notion of a limited lifetime and encouraging thoughts on valued action, seems to bear most of the observed effect on the cheating measure. We will discuss potential explanations for this in the context of the General Discussion.

This final study in the current programme of research once again provided support for the consistency and reliability in the inverse correlation between the cheating measure and *the Do-Bad/Feel-Bad* trial-type on the DF-IRAP. As such, the current thesis supports the

DF-IRAP as a potentially useful measure in the context of predicting immoral behaviour, at least in well controlled experimental setting. In the next and final chapter the empirical research presented throughout the current thesis will first be summarized and then a range of issues raised by the work will be discussed.

Chapter 7

General Discussion

General Discussion

The general aim of the current research programme was to develop a set of IRAPs that could be used to explore verbal networks related to morality and predict the occurrence of immoral behaviour. This final chapter will summarise the major findings of the six empirical studies presented in this thesis and will consider a number of conceptual, theoretical, and methodological issues arising from the work, as well as suggest new directions for further research.

Overview of the research programme

In Study 1 (presented in Chapter 4), we initiated exploratory work on capturing verbal networks related to morality by designing and implementing two IRAPs, one that targeted thoughts about good and bad actions and one that targeted frequency of moral and immoral behaviour. The ultimate goal was to determine whether the two IRAPs could be used to predict cheating behaviour in a math task that has been used for that purpose. In order to isolate difficulties with the cheating task, we performed a conceptual replication with a different cheating task in Study 2, with results that were inconclusive, but with some support for our first findings. Although the specifics will be discussed in the following section that compiles the effects for each of the IRAPs we used throughout the thesis, results from this first part of the current research programme suggested that the two IRAPs were able to predict cheating to a certain extent, and revealed that participants responded to certain trial-types according to common sense expectations, but presented interesting response patterns in others.

A secondary goal of the first part was to examine the relationships between the hypothetical process of "moral disengagement" and the IRAPs. Specifically, we expected to find a relationship between certain trial-types in the IRAPs and indices of moral

disengagement, because this social psychology concept describes how people justify certain immoral choices and actions in order to maintain a positive moral self-image. In event, however, no such relationship was found in the current research programme.

The second part of the current research programme, described in Chapter 5, consisted of two studies in which we sought to deepen our exploration of deictic responding by designing an IRAP that tapped into the feelings towards engaging in moral or immoral actions. We hypothesised performance on the IRAPs would correlate with the cheating measure and would thus be a predictor of immoral behaviour. Our literature review also pointed to a relation between willingness to engage in immoral behaviour and psychopathic traits such as boldness and impulsivity, and for that reason we decided to explore whether a well-known index of non-clinical psychopathy would correlate with performance on the IRAPs and cheating measure. In Study 3 we found the expected common-sense effects for certain trial-types in the IRAP, along with intriguing correlations between IRAP performance, cheating, and secondary psychopathy. These results, for the most part, were replicated in Study 4, in which we used the same measures in a Colombian sample in order to address the issue of ecological validity.

In the third and final part of the empirical programme we took interest in the possibility of testing the effects of a values-oriented procedure, called the Mortality Salience Induction, on cheating behaviour and the ability of the IRAP to detect these effects. In Study 5 we found a significant effect of the task when compared to a control group, and we decided to conduct another replication in a Spanish-speaking sample in Study 6, which largely reproduced our results.

This collection of findings generally demonstrated the viability of using the IRAP to assess moral behaviour, its ability to predict cheating to a certain extent, and its potential usefulness to examine the effects of interventions in the moral domain. Throughout the

entire research programme we conceived and tested three different IRAPs, each of which intended to examine a particular type of moral responding and yielded important results that we will now discuss.

IRAP effects

The CM-IRAP. This particular IRAP was conceived as a device to observe ways in which moral labels are assigned to actions in verbal networks by asking participants to categorise moral and immoral actions as good or bad. This relatively easy test was expected to yield strong effects in agreement with common-sense labelling (i.e., consider good actions as moral and bad actions and immoral), and to provide an adequate starting point for an exploratory research programme like the one presented in the current thesis. In general, throughout the studies that used the CM-IRAP, these common-sense expected effects were found for the *Good/Moral* and *Bad/Immoral* trial-types, which asked participants to respond that good actions were good and bad actions were bad. In terms of the REC model (described in Chapter 2), this implies that responding to good as moral and bad as immoral is well-established in the participants' verbal networks and is therefore a low-derivation and low-complexity type of responding.

However, the CM-IRAP also delivered paradoxical effects. For example, intuitively one might expect that people who can readily respond that moral actions are good can also quickly and strongly deny that moral actions are bad or that immoral actions are good (after all it seems to be the very same question, only asked differently). However, the CM-IRAP consistently found low effects and unclear responding patterns to the *Good/Immoral* and *Bad/Moral* trial-types, implying that there had been fewer opportunities to respond in those ways throughout the formation of the participants' verbal histories thus indicating higher levels of derivation for those trial-types.

A potential explanation for why there are fewer opportunities to derive those

relations in the natural environment is that traditional, common-sense conceptions of morality tend to be biased towards binary models: people are either good or bad, innocent or guilty, lawful or criminal. Moreover, these models generally presume that the quality of being moral or immoral is relatively stable – in fact, most of our legal systems are based on this idea, and even certain everyday expressions contain it ("once a thief, always a thief", "the leopard does not change its spots"). It is likely that, given our cultural predilection for those polar models of morality, individuals have fewer opportunities to respond to the *Good-Immoral* and *Bad-Moral* trial-types and this was reflected by performance on the CM-IRAP.

Another interesting effect that emerged from the CM-IRAP was the counter-intuitive correlation between the cheating measure and the *Bad/Immoral* trial type, indicating that people who confirmed more strongly that bad behaviour was immoral also tended to cheat more. Our interpretation of this finding is that strong, convincing affirmations that bad actions are immoral are developed throughout a verbal history as a curtain behind which immorality can more safely take place. In other words, a person who tends to lie, cheat and deceive as a functional class of behaviour may tend to lie and deceive in the context of convincing others of how aversive he or she finds immoral behaviour to be. In other words, the cheater, by definition, will frequently lie about cheating itself. In relational terms, this involves becoming highly practised at criticising immoral behaviour, as reflected in the CM-IRAP performances

As interesting and promising as these findings are, they only paint part of the picture. There is only so much knowledge to be gained about verbal networks related to morality if only these types of conceptual networks are targeted, because they tell more about the culture in which a person learned moral behaviour than the vicissitudes of individual morality. Therefore, as part of our research programme we wanted to test the ability of the IRAP to address the moral component of perspective-taking or deictic relational responding. Two types of deictic responding were explored in the current research programme by means

of two different IRAPs. In Studies 1 and 2 we used an IRAP (DM-IRAP) that asked participants to make implicit reports of the frequency of their own moral and immoral behaviour, which again provided a starting point to explore deictic responding. However, for the remainder of the current research programme, we decided to investigate how participants would feel when engaging in moral and immoral actions. We called the latter IRAP the DF (for deictic/feelings) IRAP. We will now summarise our findings using both tools.

The DM-IRAP. The deictic morality (DM) IRAP asked participants to report whether they engaged in good or bad actions frequently or rarely. Our literature review led us to expect that participants would report being often good and rarely bad, given that people generally think highly of their own morals and regard themselves as "good people" (Jordan et al., 2011); this was true to a certain extent, but the DM-IRAP also yielded some results that could be construed as counter-intuitive. We found the expected strong effects on the *Good/Often* trial-type, implying that participants considered that they frequently engaged in good actions, but we failed to find significant effects on the *Bad/Rarely* trial-type, which indicated that participants could not strongly confirm that their immoral behaviour was infrequent.

Counterintuitively, the *Bad/Often* trial-type was a predictor of immoral behaviour and revealed an unexpected anti-moral bias in both studies. The implication is that participants who more strongly denied engaging in bad behaviour also tended to cheat more, and fits together with results from the CM-IRAP to paint a picture of concealed immorality through well-practised relational responses. Indeed, these relatively strong effects on the *Good/Often* (pro-moral) and *Bad/Often* (anti-moral) trial types indicate that people seem to think that their behaviour is frequently good, but also consider it to be often bad. This is seemingly contradictory, but given that morality involves different sub-domains and

contexts, it is therefore possible to say that a person can often behave well (tells the truth, pays for his train ticket even in the absence of inspectors) but also engages in unethical behaviour frequently (tells lies, secretly reads his significant other's phone messages). These findings are in agreement to some extent with contemporary literature on moral behaviour. For example, Gino (2015) pointed out that even people who care about morality behave immorally, and do so often. In effect, morality is malleable, and people are not always able to tell when they have crossed an ethical boundary.

The DF-IRAP. For the remainder of the research programme, we decided to delve into the emotional component of moral responding by asking participants to report their feelings when engaging in moral and immoral actions. This new instrument consistently produced strong expected effects on the *Do-Good/Feel-Good* and *Do-Bad/Feel-Bad* trial-types, indicating that participants experience positive feelings when carrying out moral actions and negative feelings when engaging in immoral actions. Responses to the remaining trial-types were less clear, specifically to the *Do-Good/Feel-Bad* trial-type. In the latter case, it may be that participants have had relatively fewer opportunities throughout their verbal histories to perform this type of responding ("do you feel bad after doing something good?"). Alternatively, the lack of a significant effect might indicate that sometimes engaging in good actions does bring negative feelings (e.g., paying your taxes may be moral but not enjoyable), and thus the responses on this trial-type are not as clear due to less well-established relational networks involving this type of responding.

When considering the performance on the DF-IRAP as a predictor of cheating, we found consistent weak-to-moderate inverse correlations between the *Do-bad/Feel-bad* trial-type and the deception measure, indicating that lower probabilities of cheating were found amongst those who more strongly confirmed that negative feelings accompany bad behaviours. This contrasts with the DM-IRAP, where strong denials of engagement in

immoral behaviour predicted higher cheating. The two deictic IRAPs suggest, therefore, that under time pressure people who cheat and deceive tend to deny doing so (DM-IRAP) and confirm less readily that they feel bad when they cheat (DF-IRAP).

Up to this point, therefore, two different sets of relational networks, which are revealed under time pressure, seem to be involved in acts of dishonesty or cheating. A possible explanation for the inverse correlation between actual cheating and denial of cheating (on the DM-IRAP) could be that both behaviours overlap functionally. That is, people who tend to cheat will probably deny doing so more strongly, precisely because that denial is in itself a form of cheating. A possible explanation for the positive correlation between lower levels of cheating and confirmation of feeling bad when engaging in immoral actions seems more obvious. That is, one would expect lower levels of a particular behaviour if that behaviour evokes aversive consequences (in this case negative feelings about the self). Although tentative, these findings suggest that efforts to reduce immoral behaviour would be best focused on attempting to increase negative self-evaluation when such behaviour occurs rather than focusing on the more abstract features of immorality (e.g., the cost to the economy, etc). Of course, the current findings are largely correlational and thus point to behaviour-behaviour relations, rather than to contextual variables, which might be manipulated in order to influence cheating behaviours directly. The latter part of the research programme focused more on manipulable variables.

Effects of the intervention on the IRAPs. The Mortality Salience Induction tended to decrease the frequency of cheating in the MMT and produced stronger pro-moral effects on the DF-IRAP, particularly on the Do-Good/Feel-Good trial-type. In other words, increased death awareness strengthened implicit pro-moral responses and decreased the likelihood of cheating (although the latter involved a non-significant trend in the data). Interestingly, these effects seemed to depend on how mortality awareness was made salient. Specifically, the

complete MSI employed in Study 5 included a "destination" component (this *will* end – how do you feel about it?) and a "process" component (this *is* ending – what are you going to do?), corresponding to the questions and dots tasks, respectively. Although the complete MSI "worked," in Study 6 when only the dots task was used a similar impact on the IRAP effects and cheating was observed. Thus, it appears that temporarily increasing the salience of death as a process (i.e., towards which everyone is moving), is sufficient to increase pro-moral implicit biases, and perhaps offer a "protective" factor against choices to engage in immoral behaviour.

In this sense, the DF-IRAP proved to be useful in testing the effects of an intervention component. That is, it was able to discriminate the effects of the MSI relative to controls in Study 5 and of the dots sub-task relative to the questions sub-task in Study 6. These results are encouraging and promising because they establish the DF-IRAP as a viable option to examine deictic moral relational responding within a behavioural framework of morality.

Moral disengagement, psychopathy and IRAP performance

So far, our interpretation operates within a contextualistic, functional, clearly behavioural account of morality, but it seems useful to establish links between our findings and more traditional perspectives in Psychology. Indeed, some of the instruments we used aimed to cross into domains related to moral psychology but addressed from other theoretical and methodological perspectives. In the first part of our research programme we explored relations between the IRAPs, the cheating measure and moral disengagement, and later on we also investigated the role of psychopathy. Interestingly, the performances on DF-IRAP failed to correlate with the moral disengagement scales and also with primary psychopathy, but it did correlate with secondary psychopathy. Explaining this pattern of results requires looking at the scales themselves.

When one examines the types of questions that are asked in the moral disengagement scale and the primary psychopathy subscale, they could be seen as more likely to evoke responses that involve self-presentation biases than the secondary psychopathy subscale. Consider, for example, the following small selection of items from the moral disengagement scales:

- "Some people deserve to be treated like animals" (MDS item 7)
- "Someone who is obnoxious does not deserve to be treated like a human being" (MDS item 23)
- "Rivals deserve being humiliated and maltreated" (CMD item 28).
- "Using force is often inevitable to protect one's own interests" (CMD item 22).

The wording of these statements makes it likely that ordinary respondents will answer in a socially desirable way, indicating some level of disagreement with such sentiments. This is likely even despite reassurance that responses are confidential, anonymous and will not have any consequences on their daily lives.

The same may be true of the primary psychopathy sub-factor of the Levenson Psychopathy Scale. For instance, consider items such as: "*I enjoy manipulating other people's feelings*" or "*Success is based on survival of the fittest – I am not concerned about the losers*" in contrast to sentences from the secondary psychopathy factor, such as "*I am often bored*" or "*I quickly lose interests in tasks I start*". In the first two cases, it is relatively easy to appreciate that self-presentation biases may be involved, but in the latter case less so. Clearly, many of us would not wish to be seen as someone who does not readily and willingly go to the aid of someone in distress and certainly many of us would not like to be seen as ruthlessly self-serving with scant regard for the feelings and welfare of others. In contrast, admitting that one is easily bored does not necessarily imply anything negative about the self. Insofar as the IRAP is largely uncontaminated by self-presentational biases, it makes sense that it

correlated with only those questions with less potential for self-presentational responses (i.e., those related to secondary psychopathy). Of course this is a post-hoc interpretation of the current findings but it does provide an interesting basis for future research, something we will return to later.

Strengths, limitations, and new directions for research

Even though the original formulation of RFT includes some work on the development of verbal relations that underlie moral behaviour (for example Hayes & Hayes, 1994), to our knowledge this is the first programme of research that explores the study of moral behaviour, focusing on cheating in particular, from an RFT perspective. Furthermore, the current work contributes towards a very small body of existing research that has studied cheating using measures of implicit cognition – indeed, the only similar study that we were able to find in our literature review was by Perugini and Leone (2009), in which a moral self-concept IAT marginally predicted deceptive reports of a dice roll. In contrast, the current thesis presents a more complete research programme that used a well-known, standardised operationalization of cheating and three IRAPs that targeted relations between actions and concepts of morality, reports of frequency of moral behaviour, and personal feelings towards engaging in moral actions. In the end, we managed to develop a set of IRAPs that show promise for the assessment of cheating behaviour and perhaps in the long run, the prediction of immoral behaviour in general.

Our intervention component is also worth mentioning as a strong point of the present work. A few studies have examined mortality salience effects on moral judgment of transgressions (covered by Burke et al., 2010), but to our knowledge, the current research programme is the first to suggest effects of the Mortality Salience Induction on an implicit measure of cheating responses (at a significant level) and actual deceptive behaviour (non-significant trend) under controlled conditions, as a way to reduce cheating.

We would also highlight the fact that we ran experiments in two different cultural settings, using two separate languages, and found similar effects, which increases the ecological validity of our results. As we discussed in the introduction to the current thesis, one of the main concerns with traditional theories of morality in Psychology is that they are likely to be rather biased towards European and North American populations, and this has been supported by the fact that assessment tools deliver counterintuitive results when used in other populations (Snarey, 1985).

Despite the aforementioned advantages of our research programme, we must also mention some points that future research needs to address in order to gain a clearer perspective on the subject of cheating and moral behaviour. The first criticism of our work applies to a good number of studies in psychology, and it is related to the samples used. Participants in the current research programme were college students from Ireland and Colombia, and a concern has been raised numerous times (see Henrich et al., 2010) that college student samples are probably not representative of the general population in many domains, and thus conclusions extracted from this type of research might not easily generalise.

A second possible criticism applies to the IRAP in general. O'Shea, Watson and Brown (2015) recently claimed to have found a positivity bias in the IRAP that stems from the well-known finding that people tend to frame events in ways that highlight increases rather than decreases, or positivity rather than negativity. For example, it is much more likely that people will make statements such as "he is thinner" or "the boy is getting taller" or "the river is getting stronger" instead of "he is less fat" or "the boy is getting less short", or "the river is getting less weak". The presence of a general positivity bias in the current studies on cheating may help to explain that the largest IRAP effects tended to be for positive-positive trial types such as "Good-Moral", "Good-Often" and "Do-good/Feel-good", with more variability in the

other trial-types. On balance, it is important to note that it was the "negative" trial-types (Bad-Immoral and Do-bad/Feel-bad) that predicted cheating throughout our research programme, and it seems, therefore, that the validity of the IRAP did not appear to be threatened by the possible presence of a positivity bias. Indeed, the presence of a positivity bias has been noted and discussed in a number of previous articles on the IRAP (e.g. Barnes-Holmes, Murphy, et al., 2010, pp. 75-76; Scanlon, McEnteggart, Barnes-Holmes, & Barnes-Holmes, 2014). Indeed, it may even be the case that sensitivity to such positivity biases in the IRAP serves to increase its predictive validity (see Bast, Barnes-Holmes, & Barnes-Holmes, n.d.). Nevertheless, we acknowledge that the potential presence and impact of a so-called positivity bias in the IRAP should be explored in future research.

In closing, it is important to see the research presented in the current thesis as merely a starting point for further research in the areas of cheating behaviour and implicit cognition, and as such, many questions still remain. For example, would other measures of implicit cognition (or BIRRs) predict cheating behaviours more accurately than the IRAP? Would other types of cheating behaviours in the laboratory be predicted by the IRAP? And perhaps the most critical question is whether the IRAP could predict "real-world" cheating behaviours in naturalistic settings - for example, if it could discriminate between participants with a history of repeated deception and cheating behaviours, such as criminals who engage in "confidence scams", versus individuals of high moral standing. All of these and related questions remain to be answered in future research.

References

Albo, M. J., Winther, G., Tuni, C., Toft, S., & Bilde, T. (2011). Worthless donations: male deception and female counter play in a nuptial gift-giving spider. *BMC Evolutionary Biology*, *11*, 329. http://doi.org/10.1186/1471-2148-11-329

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*, 1423–1440. http://doi.org/10.1037//0022-3514.83.6.1423

Aquino, K., Reed, A., Thau, S., & Freeman, D. (2007). A grotesque and dark beauty: How moral identity and mechanisms of moral disengagement influence cognitive and emotional reactions to war. *Journal of Experimental Social Psychology*, *43*, 385–392. http://doi.org/10.1016/j.jesp.2006.05.013

Ardila, R. (2014). Filogénesis y ontogénesis de la moral. *Revista de La Academia Colombiana de Ciencias Exactas, Físicas Y Naturales*, *38*(supl.), 205–215.

Arndt, J., Schimel, J., & Goldenberg, J. L. (2003). Death can be good for your health: Fitness intentions as a proximal and distal defense against mortality salience. *Journal of Applied Social Psychology*, *33*, 1726–1746. http://doi.org/10.1111/j.1559-1816.2003.tb01972.x

Ayres, K., Conner, M., Prestwich, A., & Smith, P. (2012). Do implicit measures of attitudes incrementally predict snacking behaviour over explicit affect-related measures? *Appetite*, *58*(3), 835–841. http://doi.org/10.1016/j.appet.2012.01.019

Bandura, A. (2002). Selective Moral Disengagement in the Exercise of Moral Agency. *Journal of Moral Education*, *31*, 101–119. http://doi.org/10.1080/0305724022014322

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology*, *71*, 364–374. http://doi.org/10.1037/0022-3514.71.2.364

Barnes, C. M., Schaubroeck, J., Huth, M., & Ghumman, S. (2011). Lack of sleep and unethical conduct. *Organizational Behavior and Human Decision Processes*, *115*(2), 169–180. http://doi.org/10.1016/j.obhdp.2011.01.009

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist*, *32*, 169–177.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A Sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) Model. *Psychological Record*, *60*(3), 527–542.

Barnes-Holmes, D., Hayden, E., Barnes-Holmes, Y., & Stewart, I. (2008). The Implicit Relational Assessment Procedure (IRAP) as a response-time and event-related-potentials methodology for testing natural verbal relations: A preliminary study. *The Psychological Record*, *58*, 497–516.

Barnes-Holmes, D., Murphy, A., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). The implicit relational assessment procedure: Exploring the impact of private versus public contexts and the response latency criterion on pro-white and anti-black stereotyping among white Irish individuals. *Psychological Record*, *60*(3), 57–66.

Barnes-Holmes, D., Murtagh, L., Barnes-Holmes, Y., Stewart, I., Murphy, A., & Boles, S. (2010). Using the Implicit Association Test and the Implicit Relational Assessment Procedure to measure attitudes toward meat and vegetables in vegetarians and meat-eaters. *Psychological Record*, *60*(3), 287–306.

Bast, D. F., Barnes-Holmes, Y., & Barnes-Holmes, D. (n.d.). *Developing an Individualized Implicit Relational Assessment Procedure (IRAP) as a Potential Measure of Self-Forgiveness related to Negative and Positive Behaviour*.

Becker, E. (1973). *The denial of death*. New York: Free Press.

Bekkers, R., & Wiepking, P. (2010). Accuracy of self-reports on donations to charitable organizations. *Quality & Quantity*, *45*(6), 1369–1383. http://doi.org/10.1007/s11135-010-9341-9

Beretvas, S. N., Meyers, J. L., & Leite, W. L. (2002). A Reliability Generalization Study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement*, *62*(4), 570–589. http://doi.org/10.1177/0013164402062004003

Bevan, A. L., Maxfield, M., & Bultmann, M. N. (2014). The effects of age and death awareness on intentions for healthy behaviours. *Psychology & Health*, *29*(February 2015), 405–421. http://doi.org/10.1080/08870446.2013.859258

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*(5), 760–773. http://doi.org/10.1037//0022-3514.81.5.760

Brewis, A. A., & Wutich, A. (2012). Explicit versus implicit fat-stigma. *American Journal of Human Biology*, *24*(3), 332–338. http://doi.org/10.1002/ajhb.22233

British Psychological Society. (2009). *Code of Ethics and Conduct: Guidance published by the Ethics Committee of the British Psychological Society*. Leicester: British Psychological Society.

Brown, C., Garwood, M. P., & Williamson, J. E. (2012). It pays to cheat: tactical deception in a cephalopod social signalling system. *Biology Letters*, *8*(5), 729–32. http://doi.org/10.1098/rsbl.2012.0435

Brown, R. P., Budzek, K., & Tamborski, M. (2009). On the Meaning and Measure of Narcissism. *Personality and Social Psychology Bulletin*, *35*, 951–964. http://doi.org/10.1177/0146167209335461

Brown, R. P., Tamborski, M., Wang, X., Barnes, C. D., Mumford, M. D., Connelly, S., & Devenport, L. D. (2011). Moral Credentialing and the Rationalization of Misconduct. *Ethics & Behavior*, *21*, 1–12. http://doi.org/10.1080/10508422.2011.537566

Burke, B. L., Martens, A., & Faucher, E. H. (2010). Two decades of terror management theory: a meta-analysis of mortality salience research. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*, *14*, 155–195. http://doi.org/10.1177/1088868309352321

Burman, E. (1999). Morality and the goals of development. In M. Woodhead, D. Faulkner, & K. Littleton (Eds.), *Making sense of social development* (1st ed., pp. 170–180). London: Routledge.

Cai, F., & Wyer, R. S. (2014). The impact of mortality salience on the relative effectiveness of donation appeals. *Journal of Consumer Psychology*, *25*(1), 101–112. http://doi.org/10.1016/j.jcps.2014.05.005

Cai, H., Sriram, N., Greenwald, A. G., & McFarland, S. G. (2004). The Implicit Association Test's D Measure Can Minimize a Cognitive Skill Confound: Comment on McFarland and Crouch (2002). *Social Cognition*, *22*(6), 673–684. http://doi.org/10.1521/soco.22.6.673.54821

Callegaro Borsa, J., Figuereido Damásio, B., & Ruschel Bandeira, D. (2012). Cross-cultural adaptation and validation of psychological instruments: some considerations. *Paidéia (Ribeirão Preto)*, *22*(53), 423–432.

Campos Vizcarra, S. (2013). *El manejo del terror y su impacto en los motivos de la identidad: un estudio experimental*. Pontificia Universidad Católica del Perú.

Caprara, G. V., Fida, R., Vecchione, M., Tramontano, C., & Barbaranelli, C. (2009). Assessing civic moral disengagement: Dimensionality and construct validity. *Personality and Individual Differences*, *47*, 504–509. http://doi.org/10.1016/j.paid.2009.04.027

Carpenter, K. M., Martinez, D., Vadhan, N. P., Barnes-Holmes, D., & Nunes, E. V. (2012). Measures of attentional bias and relational responding are associated with behavioral treatment outcome for cocaine dependence. *The American Journal of Drug and Alcohol Abuse*, *38*(2), 146–54. http://doi.org/10.3109/00952990.2011.643986

Castelli, L., Zogmaister, C., & Tomelleri, S. (2009). The transmission of racial attitudes within the family. *Developmental Psychology*, *45*(2), 586–591. http://doi.org/10.1037/a0014619

Chapman, H. a, & Anderson, A. K. (2014). Trait physical disgust is related to moral judgments outside of the purity domain. *Emotion (Washington, D.C.)*, *14*(2), 341–8. http://doi.org/10.1037/a0035120

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex (New York, N.Y. : 1991)*, *19*(7), 1557–66. http://doi.org/10.1093/cercor/bhn189

Chugh, D., Kern, M. C., Zhu, Z., & Lee, S. (2014). Withstanding moral disengagement: Attachment security as an ethical intervention. *Journal of Experimental Social Psychology*, *51*, 88–93. http://doi.org/10.1016/j.jesp.2013.11.005

Claybourn, M. (2010). Relationships Between Moral Disengagement, Work Characteristics and Workplace Harassment. *Journal of Business Ethics*, *100*(2), 283–301. http://doi.org/10.1007/s10551-010-0680-1

Congreso de la República de Colombia. Ley 1090 de 2006 (2006). Bogotá: Congreso de la República de Colombia.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. http://doi.org/10.1037/h0047358

De Houwer, J. (2001). A Structural and Process Analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, *37*(6), 443–451. http://doi.org/10.1006/jesp.2000.1464

De Houwer, J. (2003). The Extrinsic Affective Simon Task. *Experimental Psychology (formerly "Zeitschrift Für Experimentelle Psychologie")*, *50*(2), 77–85. http://doi.org/10.1026//1618-3169.50.2.77

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, *37*(2), 176–187. http://doi.org/10.1016/j.lmot.2005.12.002

De Houwer, J., Hermans, D., & Spruyt, A. (2001). Affective Priming of Pronunciation Responses: Effects of Target Degradation. *Journal of Experimental Social Psychology*, *37*(1), 85–91. http://doi.org/10.1006/jesp.2000.1437

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368. http://doi.org/10.1037/a0014211

De Waal, F. B. M., Smith Churchland, P., Pievani, T., & Parmigiani, S. (2014). Evolved morality: The biology and philosophy of human conscience. *Behaviour*, *151*, 137–141.

Dean, A. C., Altstein, L. L., Berman, M. E., Constans, J. I., Sugar, C. A., & McCloskey, M. S. (2013). Secondary Psychopathy, but not Primary Psychopathy, is Associated with Risky Decision-Making in Noninstitutionalized Young Adults. *Personality and Individual Differences*, *54*(2), 272–277.

Devany, J. M., Hayes, S. C., & Nelson, R. O. (1986). Equivalence class formation in language-able and language-disabled children. *Journal of the Experimental Analysis of Behavior*, *46*(3), 243–257. http://doi.org/10.1901/jeab.1986.46-243

Dugdale, N., & Lowe, C. F. (1990). Naming and stimulus equivalence. In D. E. Blackman & H. Lejeune (Eds.), *Behaviour analysis in theory and practice: Contributions and controversies* (pp. 115–138). Hillsdale, NJ: Lawrence Erlbaum.

Dymond, S. (2014). Meaning is more than associations: relational operants and the search for derived relations in nonhumans. *Journal of the Experimental Analysis of Behavior*, *101*(1), 152–155.

Dymond, S., O'Hora, D., Whelan, R., & Donovan, A. O. (2006). Citation Analysis of Skinner ' s Verbal Behavior : 1984 – 2004. *The Behavior Analyst*, *1*(1), 75–88.

Eckhardt, C. I., Samper, R., Suhr, L., & Holtzworth-Munroe, A. (2012). Implicit Attitudes Toward Violence Among Male Perpetrators of Intimate Partner Violence: A Preliminary Investigation. *Journal of Interpersonal Violence*, *27*(3), 471–491. http://doi.org/10.1177/0886260511421677

Epstein, J., Santo, R. M., & Guillemin, F. (2014). A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *Journal of Clinical Epidemiology*, *68*(4), 435–441. http://doi.org/10.1016/j.jclinepi.2014.11.021

Falkenbach, D., Poythress, N., & Creevy, C. (2008). The exploration of subclinical psychopathic subtypes and the relationship with types of aggression. *Personality and Individual Differences*, *44*(4), 821–832. http://doi.org/10.1016/j.paid.2007.10.012

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: the relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–41. http://doi.org/10.1016/j.cognition.2012.02.001

Fischbacher, U., & Heusi, F. (2008). Lies in disguise, an experimental study on cheating. *Research Paper Series Thurgau Institute of Economics and Department of Economics at the University of Konstanz*, *40*, 1–20.

Francken, J. C., Gaal, S. van, & de Lange, F. P. (2011). Immediate and long-term priming effects are independent of prime awareness. *Consciousness and Cognition*, *20*(4), 1793–1800. http://doi.org/10.1016/j.concog.2011.04.005

Frimer, J. A., & Walker, L. J. (2008). Towards a new paradigm of moral personhood. *Journal of Moral Education*, *37*, 333–356. http://doi.org/10.1080/03057240802227494

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21–38.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, Mass: Harvard University Press.

Gino, F. (2015). Understanding ordinary unethical behavior: why people who value morality act immorally. *Current Opinion in Behavioral Sciences*, *3*(im), 107–111. http://doi.org/10.1016/j.cobeha.2015.03.001

Gino, F., & Mogilner, C. (2014). Time, money, and morality. *Psychological Science*, *25*(2), 414–21. http://doi.org/10.1177/0956797613506438

Gino, F., Schweitzer, M. E., Mead, N. L., & Ariely, D. (2011). Unable to resist temptation: How self-control depletion promotes unethical behavior. *Organizational Behavior and Human Decision Processes*, *115*(2), 191–203. http://doi.org/10.1016/j.obhdp.2011.03.001

Gonzalez-Ocantos, E., de Jonge, C. K., Meléndez, C., Osorio, J., & Nickerson, D. W. (2012). Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua. *American*

Journal of Political Science, *56*(1), 202–217. http://doi.org/10.1111/j.1540-5907.2011.00540.x

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046. http://doi.org/10.1037/a0015141

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*, 366–385. http://doi.org/10.1037/a0021847

Greenberg, J., Pyszczynski, T., Solomon, S., Simon, L., & Breus, M. (1994). Role of Consciousness and Accessibility of Death-Related Thoughts in Mortality Salience Effects. *Journal of Personality and Social Psychology*, *67*(4), 627–637. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0028520229&partnerID=tZOtx3y1

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Soclal Psychology*, *74*(6), 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216.

Haider, A. H., Sexton, J., Sriram, N., Cooper, L. A., Efron, D. T., Swoboda, S., … Cornwell, E. E. (2011). Association of Unconscious Race and Social Class Bias With Vignette-Based Clinical Assessments by Medical Students. *JAMA: The Journal of the American Medical Association*, *306*(9), 942–951. http://doi.org/10.1001/jama.2011.1248

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, *108*(4), 814–834. http://doi.org/10.1037//0033-295X.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*, 55–66. http://doi.org/10.1162/0011526042365555

Hauck-Filho, N., & Teixeira, M. A. P. (2014). Revisiting the psychometric properties of the Levenson self-report psychopathy scale. *Journal of Personality Assessment*, *96*(4), 459–64. http://doi.org/10.1080/00223891.2013.865196

Hayes, L. J., Adams, M. A., & Rydeen, K. L. (1994). Ethics, Choice and Value. In L. J. Hayes, G. J. Hayes, S. L. Moore, & P. M. Ghezzi (Eds.), *Ethical Issues in Developmental Disabilities* (pp. 11–39). Reno, NV: Context.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory : a post-Skinnerian account of human language and cognition*. New York: Kluwer Academic/Plenum Publishers.

Hayes, S. C., Gifford, E. V., & Hayes, G. J. (1998). Moral Behavior and the Development of Verbal Regulation. *The Behavior Analyst*, *21*(2), 253–279.

Hayes, S. C., & Hayes, G. J. (1994). Stages of Moral Development as Stages of Rule-Governance. In L. J. Hayes, G. J. Hayes, S. C. Moore, & P. M. Ghezzi (Eds.), *Ethical Issues in Developmental Disabilities* (pp. 45–65). Reno, NV: Context.

Henrich, J., Heine, S. J., & Noren. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83. Retrieved from http://www2.psych.ubc.ca/~henrich/pdfs/WeirdPeople.pdf

Hilbig, B., & Hessler, C. (2012). What lies beneath: How the distance between truth and lie drives dishonesty. *Journal of Experimental Social Psychology*, *49*(2), 263–266. http://doi.org/10.1016/j.jesp.2012.11.010

Holden, R. R. (2007). Socially desirable responding does moderate personality scale validity both in experimental and in nonexperimental contexts. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *39*(3), 184–201. http://doi.org/10.1037/cjbs2007015

Horner, A. J., & Henson, R. N. (2008). Priming, response learning and repetition suppression. *Neuropsychologia*, *46*(7), 1979–1991. http://doi.org/10.1016/j.neuropsychologia.2008.01.018

Hughes, S., & Barnes-Holmes, D. (2014). Associative concept learning, stimulus equivalence, and relational frame theory: working out the similarities and differences between human and nonhuman behavior. *Journal of the Experimental Analysis of Behavior*, *101*(1), 156–160. http://doi.org/10.1002/jeab.60

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The Dominance of Associative Theorizing in Implicit Attitude Research: Propositional and Behavioral Alternatives. *Psychological Record*, *61*(3), 465–496. Retrieved from http://web.ebscohost.com/ehost/detail?sid=661d13ef-80e4-43f2-9681-aa3926455918%40sessionmgr12&vid=1&hid=18&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=a9h&AN=65033134

Hughes, S., Barnes-Holmes, D., & Vahey, N. (2012). Holding on to our functional roots when exploring new intellectual islands: A voyage through implicit cognition research. *Journal of Contextual Behavioral …*, *1*(1), 17–38. Retrieved from http://www.sciencedirect.com/science/article/pii/S2212144712000075

Hussey, I., & Barnes-Holmes, D. (2015). The malleability of implicit attitudes to death after a modified mortality salience induction.

Hussey, I., Barnes-Holmes, D., & Barnes-Holmes, Y. (2015). From Relational Frame Theory to implicit attitudes and back again: claryfing the link between RFT and IRAP research. *Current Opinion in Psychology*, *2*, 11–15.

Jackson, L. E., & Gaertner, L. (2010). Mechanisms of moral disengagement and their differential use by right-wing authoritarianism and social dominance orientation in support of war. *Aggressive Behavior*, *36*(4), 238–250. http://doi.org/Article

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science (New York, N.Y.)*, *310*(5745), 116–9. http://doi.org/10.1126/science.1111709

Johnson, M. M., & Rosenfeld, J. P. (1992). Oddball-evoked P300-based method of deception detection in the laboratory. II: Utilization of non-selective activation of relevant knowledge. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *12*, 289–306. http://doi.org/10.1016/0167-8760(92)90067-L

Jones, A., & Fitness, J. (2008). Moral hypervigilance: the influence of disgust sensitivity in the moral domain. *Emotion (Washington, D.C.)*, *8*(5), 613–627. http://doi.org/10.1037/a0013435

Jones, T. M. (1991). Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model. *The Academy of Management Review*, *16*, 366. http://doi.org/10.2307/258867

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the Moral Self: The Effects of Recalling Past Moral Actions on Future Moral Behavior. *Personality and Social Psychology Bulletin*, *37*, 701–713. http://doi.org/10.1177/0146167211400208

Kammeyer-Mueller, J. D., Simon, L. S., & Rich, B. L. (2012). The Psychic Cost of Doing Wrong: Ethical Conflict, Divestiture Socialization, and Emotional Exhaustion. *Journal of Management*, *38*(3), 784–808. http://doi.org/10.1177/0149206310381133

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., & Fessler, D. M. T. (2007). Harm, Affect, and the Moral/Conventional Distinction. *Mind & Language*, *22*(2), 117–131. http://doi.org/10.1111/j.1468-0017.2007.00302.x

Klein, C. (2011). The Dual Track Theory of Moral Decision-Making: a Critique of the Neuroimaging Evidence. *Neuroethics*, *4*, 143–162. http://doi.org/10.1007/s12152-010-9077-1

Kohlberg, L., & Hersh, R. H. (1977). Moral Development: A Review of the Theory. *Theory into Practice*, *16*(2), 53–59. Retrieved from http://www.jstor.org/stable/pdfplus/1475172.pdf?acceptTC=true

Kouchaki, M., & Smith, I. H. (2013). The Morning Morality Effect: The Influence of Time of Day on Unethical Behavior. *Psychological Science*, (October). http://doi.org/10.1177/0956797613498099

Krahé, B., Becker, J., & Zöllter, J. (2008). Contextual cues as a source of response bias in personality questionnaires: The case of the NEO-FFI. *European Journal of Personality*, *22*(8), 655–673. http://doi.org/10.1002/per.695

Krebs, D. L. (2008). Morality: An Evolutionary Account. *Perspectives on Psychological Science*, *3*(3), 149–172. http://doi.org/10.1111/j.1745-6924.2008.00072.x

La Vigne, N., & Lowry, S. (2011). *Evaluation of camera use to prevent crime in commuter parking facilities: a randomized controlled trial*. *Technical Report of The Urban Institute*

Justice Policy Center. Washington, D.C. Retrieved from http://www.urban.org/UploadedPDF/412451-Evaluation-of-Camera-Use-to-Prevent-Crime-in-Commuter-Parking-Facilities.pdf

Le Roux, A., Snyder-Mackler, N., Roberts, E. K., Beehner, J. C., & Bergman, T. J. (2013). Evidence for tactical concealment in a wild primate. *Nature Communications*, *4*, 1462. http://doi.org/http://dx.doi.org/10.1038/ncomms2468

Leong, F. T. L., & Lyons, B. (2010). Ethical Challenges for Cross-Cultural Research Conducted by Psychologists From the United States. *Ethics & Behavior*, *20*, 250–264. http://doi.org/10.1080/10508421003798984

Levenson, M. R., Kiehl, K. a, & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology*, *68*(1), 151–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7861311

Levine, C., Kohlberg, L., & Hewer, A. (1985). The Current Formulation of Kohlberg's Theory and a Response to Critics. *Human Development*, *28*(2), 94–100. http://doi.org/10.1159/000272945

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The Scientific Status of Projective Techniques. *Psychological Science in the Public Interest*, *1*(2), 27–66.

Morling, B., & Lamoreaux, M. (2013). Measuring Culture Outside the Head: A Meta-Analysis of Individualism—Collectivism in Cultural Products. *Personality and Social Psychology Review*, *12*(3), 199–221.

Muraven, M., Pogarsky, G., & Shmueli, D. (2006). Self-Control Depletion and the General Theory of Crime. *Journal of Quantitative Criminology*, *22*(3), 263–277. Retrieved from https://www.google.ie/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CDMQFjAA&url=http://www.albany.edu/~muraven/publications/promotion files/articles/muraven, pogarsky, & shmueli, 2006.pdf&ei=TUgiUcClFIiXhQeD8oGADA&usg=AFQjCNG1VR__JcF2OQvZdhOKNZLDVP4BuA&bvm=bv.42553238,d.ZG4

Nabi, R. L. (2002). The theoretical versus the lay meaning of disgust: Implications for emotion research. *Cognition & Emotion*, *16*(January 2015), 695–703. http://doi.org/10.1080/02699930143000437

National University of Ireland - Maynooth - Department of Psychology. (2015). Guidelines for Safe Work Practice at the Department of Psychology. Maynooth, Ireland: National University of Ireland, Maynooth.

Nicholson, E., & Barnes-Holmes, D. (2012a). Developing an implicit measure of disgust propensity and disgust sensitivity: Examining the role of implicit disgust propensity and sensitivity in obsessive-compulsive tendencies. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(3), 922–930.

Nicholson, E., & Barnes-Holmes, D. (2012b). The Implicit Relational Assessment Procedure (IRAP) as a measure of spider fear. *The Psychological Record*, *62*(2), 263–277.

Niesta, D., Fritsche, I., & Jonas, E. (2008). Mortality salience and its effects on peace processes: A review. *Social Psychology*, *39*, 48–58. http://doi.org/10.1027/1864-9335.39.1.48

Nisbett, R. E., & Wilson, T. D. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, *84*(3), 231–259.

Noordraven, E., & Verschuere, B. (2013). Predicting the Sensitivity of the Reaction Time-based Concealed Information Test. *Applied Cognitive Psychology*, *27*(3), 328–335. http://doi.org/10.1002/acp.2910

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go Association Task. *Social Cognition*, *19*(6), 625–664. Retrieved from http://www.buyologyinc.com/latency_white_paper.pdf

Ong, A. D., & Weiss, D. J. (2000). The Impact of Anonymity on Responses to Sensitive Questions. *Journal of Applied Social Psychology*, *30*, 1691–1708. http://doi.org/10.1111/j.1559-1816.2000.tb02462.x

Onions, C. T., Friedrichsen, G. W. S., & Burchfield, R. W. (1978). *The Oxford Dictionary of English Etymology*. Oxford: Clarendon.

Ortu, D. (2012). Neuroscientific measures of covert behavior. *The Behavior Analyst / MABA*, *35*(1), 75–87. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3359857&tool=pmcentrez&rendertype=abstract

Parling, T., Cernvall, M., Stewart, I., Barnes-Holmes, D., & Ghaderi, A. (2012). Using the implicit relational assessment procedure to compare implicit pro-thin/anti-fat attitudes of patients with anorexia nervosa and non-clinical controls. *Eating Disorders*, *20*(2), 127–43. http://doi.org/10.1080/10640266.2012.654056

Pelton, J., Gound, M., Forehand, R., & Brody, G. (2004). The Moral Disengagement Scale: Extension with an American Minority Sample. *Journal of Psychopathology and Behavioral Assessment*, *26*, 31–39. http://doi.org/10.1023/B:JOBA.0000007454.34707.a5

Perugini, M., & Leone, L. (2009). Implicit self-concept and moral action. *Journal of Research in Personality*, *43*(5), 747–754. http://doi.org/10.1016/j.jrp.2009.03.015

Piqueras-Fiszman, B., Velasco, C., & Spence, C. (2012). Exploring implicit and explicit crossmodal colour–flavour correspondences in product packaging. *Food Quality and Preference*, *25*(2), 148–155. http://doi.org/10.1016/j.foodqual.2012.02.010

Power, P., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). The Implicit Relational Assessment Procedure (IRAP) as a measure of implicit relative preferences: A first study. *The Psychological Record*, *59*, 621–640.

Redondo, N. (2012). *Eficacia de un programa de tratamiento psicológico para maltratadores*. Universidad Complutense de Madrid. Retrieved from http://eprints.ucm.es/15003/

Reynolds, S. J. (2006). A Neurocognitive Model of the Ethical Decision-Making Process: Implications for Study and Practice. *Journal of Applied Psychology*, *91*(4), 737–748. http://doi.org/10.1037/0021-9010.91.4.737

Rezaei, A. R. (2011). Validity and reliability of the IAT: Measuring gender and ethnic stereotypes. *Computers in Human Behavior*, *27*(5), 1937–1941. http://doi.org/10.1016/j.chb.2011.04.018

Roddy, S., Stewart, I., & Barnes-Holmes, D. (2011). Facial reactions reveal that slim is good but fat is not bad: Implicit and explicit measures of body-size bias. *European Journal of Social Psychology*, *41*(6), 688–694. http://doi.org/10.1002/ejsp.839

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523–534. http://doi.org/10.1016/j.labeco.2009.04.005

Rosenfeld, J. P., Ellwanger, J., & Sweet, J. (1995). Detecting simulated amnesia with event-related brain potentials. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *19*, 1–11. http://doi.org/10.1016/0167-8760(94)00057-L

Rosenfeld, J. P., Labkovsky, E., Winograd, M., Lui, M. a., Vandenboom, C., & Chedid, E. (2008). The Complex Trial Protocol (CTP): A new, countermeasure-resistant, accurate, P300-based method for detection of concealed information. *Psychophysiology*, *45*, 906–919. http://doi.org/10.1111/j.1469-8986.2008.00708.x

Rosenfeld, J. P., Reinhart, A. M., Bhatt, M., Ellwanger, J., Gora, K., Sekera, M., & Sweet, J. (1998). P300 correlates of simulated malingered amnesia in a matching-to-sample task: Topographic analyses of deception versus truthtelling responses. *International Journal of Psychophysiology*, *28*, 233–247. http://doi.org/10.1016/S0167-8760(97)00084-6

Routledge, C., Ostafin, B., Juhl, J., Sedikides, C., Cathey, C., & Liao, J. (2010). Adjusting to death: the effects of mortality salience and self-esteem on psychological well-being, growth motivation, and maladaptive behavior. *Journal of Personality and Social Psychology*, *99*(6), 897–916. http://doi.org/10.1037/a0021431

Sagarin, B. J., Rhoads, K. V. L., & Cialdini, R. B. (1998). Deceiver's Distrust: Denigration as a Consequence of Undiscovered Deception. *Personality and Social Psychology Bulletin*, *24*, 1167–1176. http://doi.org/10.1177/01461672982411004

Scanlon, G., McEnteggart, C., Barnes-Holmes, Y., & Barnes-Holmes, D. (2014). Using the implicit relational assessment procedure (IRAP) to assess implicit gender bias and self-esteem in typically- developing children and children with adhd and with dyslexia. *Behavioral Development Bulletin*, *19*(2), 48–59.

Scherer, L., & Lambert, A. J. (2009). Counterstereotypic Exemplars in Context: Evidence for Intracategory Differentiation using Implicit Measures. *Social Cognition*, *27*(4), 522–549. http://doi.org/10.1521/soco.2009.27.4.522

Schindler, S., Reinhard, M. A., & Stahlberg, D. (2013). Tit for tat in the face of death: The effect of mortality salience on reciprocal behavior. *Journal of Experimental Social Psychology*, *49*(1), 87–92. http://doi.org/10.1016/j.jesp.2012.06.002

Schmidt, F., Haberkamp, A., & Schmidt, T. (2011). Dos and don'ts in response priming research. *Advances in Cognitive Psychology*, *7*, 120–131. http://doi.org/10.2478/v10053-008-0092-2

Scott, R. (2014). Psychopathy – An Evolving and Controversial Construct. *Psychiatry, Psychology and Law*, (August), 1–29. http://doi.org/10.1080/13218719.2014.911056

Seymour, T. L., Seifert, C. M., Shafto, M. G., & Mosmann, A. L. (2000). Using response time measures to assess "guilty knowledge." *2Journal of Applied Psychology*, *85*(1), 30–37.

Shariff, A. F., & Norenzayan, A. (2011). Mean Gods Make Good People: Different Views of God Predict Cheating Behavior. *International Journal for the Psychology of Religion*, *21*, 85–96. http://doi.org/10.1080/10508619.2011.556990

Shea, B. O., Watson, D. G., Brown, G. D. A., Shea, B. O., Watson, D. G., & Brown, G. D. A. (2015). Psychological Assessment Measuring Implicit Attitudes : A Positive Framing Bias Measuring Implicit Attitudes : A Positive Framing Bias Flaw in the Implicit Relational Assessment Procedure (IRAP). *Psychological Assessment*. http://doi.org/http://dx.doi.org/10.1037/pas0000172

Shu, L. L., Gino, F., & Bazerman, M. H. (2011). Dishonest Deed, Clear Conscience: When Cheating Leads to Moral Disengagement and Motivated Forgetting. *Personality and Social Psychology Bulletin*, *37*, 330–349. http://doi.org/10.1177/0146167211398138

Shweder, R. A., Mahapatra, M., & Miller, J. G. (1987). Culture and Moral Development. In J. Kagan & S. Lamb (Eds.), (pp. 1–82). Chicago: University of Chicago Press.

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech and Hearing Research*, *14*(1), 5–13.

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Boston, MA: Authors Cooperative.

Skinner, B. F. (1957). *Verbal Behavior*. Prentice-Hall.

Skinner, B. F. (1966). An operant analysis of problem solving. In B. Kleinmutz (Ed.), *Problem solving: Research, method, and theory* (pp. 225–257). New York, New York, USA: John Wiley & Sons.

Skinner, B. F. (1971). *Beyond freedom & dignity*. Indianapolis Ind.: Hackett Pub.

Skinner, B. F. (1975). The ethics of helping people. *Criminal Law Bulletin, 11*, 623–636. http://doi.org/10.1037/a0009052

Snarey, J. R. (1985). Cross-cultural universality of social-moral development: A critical review of Kohlbergian research. *Psychological Bulletin*, *97*(2), 202–232. http://doi.org/10.1037/0033-2909.97.2.202

Sobhani, M., & Bechara, A. (2011). A somatic marker perspective of immoral and corrupt behavior. *Social Neuroscience*, *6*(5-6), 640–52. http://doi.org/10.1080/17470919.2011.605592

Solomon, S., Greenberg, J., & Pyszczynski, T. (1991). *A terror management theory of social behavior: The psychological functions of self-esteem and cultural worldviews*. *Advances in Experimental Social Psychology* (Vol. 24). Elsevier. http://doi.org/10.1016/S0065-2601(08)60328-7

Soreth, M. (2011). The False Dichotomy of Morality and Self-Interest as Determinants of Action: Facilitating Intervention against Genocide. *Behavior and Social Issues*, *20*, 32–43. http://doi.org/10.5210/bsi.v20i0.2468

Spence, S. A., Farrow, T. F., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. (2001). Behavioural and functional anatomical correlates of deception in humans. *Neuroreport*, *12*(13), 2849–53. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11588589

Spence, S. A., & Kaylor-Hughes, C. J. (2008). Looking for truth and finding lies: the prospects for a nascent neuroimaging of deception. *Neurocase*, *14*(1), 68–81. http://doi.org/10.1080/13554790801992776

Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental Psychology (formerly "Zeitschrift Für Experimentelle Psychologie")*, *56*(4), 283–294. http://doi.org/10.1027/1618-3169.56.4.283

Stanley, D., Phelps, E., & Banaji, M. R. (2008). The Neural Basis of Implicit Attitudes. *Current Directions in Psychological Science*, *17*(2), 164–170.

Stock, A., & Stock, C. (2004). A short history of ideo-motor action. *Psychological Research*, *68*(2-3), 176–188. http://doi.org/10.1007/s00426-003-0154-5

Suhler, C. L., & Churchland, P. (2011). Can Innate, Modular "Foundations" Explain Morality? Challenges for Haidt's Moral Foundations Theory. *Journal of Cognitive Neuroscience*, *23*, 2103–2116. http://doi.org/10.1162/jocn.2011.21637

Sundberg, M. L., Partington, J. W., & J.W., P. (1998). *Teaching language to children with autism or other developmental disabilities*. Danville, CA: Behavior Analysts Inc.

Takahashi, H., Kato, M., Matsuura, M., Koeda, M., Yahata, N., Suhara, T., & Okubo, Y. (2008). Neural Correlates of Human Virtue Judgment. *Cerebral Cortex*, *18*(8), 1886–1891. http://doi.org/10.1093/cercor/bhm214

Tassy, S., Deruelle, C., Mancini, J., Leistedt, S., & Wicker, B. (2013). High levels of psychopathic traits alters moral choice but not moral judgment. *Frontiers in Human Neuroscience*, *7*(June), 229. http://doi.org/10.3389/fnhum.2013.00229

Tenbrunsel, A. E., Diekmann, K. a., Wade-Benzoni, K. a., & Bazerman, M. H. (2010). The ethical mirage: A temporal explanation as to why we are not as ethical as we think we are. *Research in Organizational Behavior*, *30*, 153–173. http://doi.org/10.1016/j.riob.2010.08.004

Teper, R., & Inzlicht, M. (2010). Active Transgressions and Moral Elusions: Action Framing Influences Moral Behavior. *Social Psychological and Personality Science*, *2*, 284–288. http://doi.org/10.1177/1948550610389338

Tobey Klass, E. (1978). Psychological effects of immoral actions: the experimental evidence. *Psychological Bulletin*, *85*(4), 756–771. http://doi.org/10.1037/0033-2909.85.4.756

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry*, *48C*, 59–65. http://doi.org/10.1016/j.jbtep.2015.01.004

Van Bockstaele, B., Verschuere, B., Moens, T., Suchotzki, K., Debey, E., & Spruyt, A. (2012). Learning to lie: effects of practice on the cognitive cost of lying. *Frontiers in Psychology*, *3*, 526. http://doi.org/10.3389/fpsyg.2012.00526

Van Ravenzwaaij, D., van der Maas, H. L. J., & Wagenmakers, E.-J. (2011). Does the Name-Race Implicit Association Test Measure Racial Prejudice? *Experimental Psychology (formerly Zeitschrift Für Experimentelle Psychologie)*, *58*(4), 271–277. http://doi.org/10.1027/1618-3169/a000093

Vendemia, J. M. C., Buzan, R. F., & Green, E. P. (2005). Practice Effects, Workload, and Reaction Time in Deception. *The American Journal of Psychology*, *118*(3), 413–429.

Verschuere, B., Spruyt, A., Meijer, E. H., & Otgaar, H. (2011). The ease of lying. *Consciousness and Cognition*, *20*(3), 908–11. http://doi.org/10.1016/j.concog.2010.10.023

Vohs, K. D., & Schooler, J. W. (2008). The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating. *Psychological Science*, *19*, 49–54. http://doi.org/10.1111/j.1467-9280.2008.02045.x

Von Hippel, W., Lakin, J. L., & Shakarchi, R. (2005). Individual Differences in Motivated Social Cognition: The Case of Self-Serving Information Processing. *Personality and Social Psychology Bulletin*, *31*, 1347–1357. http://doi.org/10.1177/0146167205274899

Walker, L. J. (2006). Gender and Morality. In M. Killen & J. Smetana (Eds.), *Handbook of Moral Development* (pp. 93–115). Mahwah, NJ: Lawrence Erlbaum. Retrieved from https://books.google.ie/books?hl=en&lr=&id=4CV5AgAAQBAJ&oi=fnd&pg=PA93&dq=gender++and+morality&ots=0N46EV3TQa&sig=1qUmnmBLxArHT0tU3c3nZsrdbws&redir_esc=y#v=onepage&q=gender and morality&f=false

Wheeler, B. C. (2008). Selfish or altruistic? An analysis of alarm call function in wild capuchin monkeys, Cebus apella nigritus. *Animal Behaviour*, *76*(5), 1465–1475. http://doi.org/10.1016/j.anbehav.2008.06.023

Wheeler, B. C. (2009). Monkeys crying wolf? Tufted capuchin monkeys use anti-predator calls to usurp resources from conspecifics. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1669), 3013–3018. http://doi.org/10.1098/rspb.2009.0544

Williams, E. J., Bott, L. A., Patrick, J., & Lewis, M. B. (2013). Telling Lies: The Irrepressible Truth? *PLoS One*, *8*(4), e60713. http://doi.org/10.1371/journal.pone.0060713

Yang, Q., Li, A., Xiao, X., Zhang, Y., & Tian, X. (2014). Dissociation between morality and disgust: An event-related potential study. *International Journal of Psychophysiology*, *94*, 84–91. http://doi.org/10.1016/j.ijpsycho.2014.07.008

Young, L., Bechara, A., Tranel, D., Damasio, H., Hauser, M. D., & Damasio, A. (2010). Damage to Ventromedial Prefrontal Cortex Impairs Judgment of Harmful Intent. *Neuron*, *65*(6), 845–851. http://doi.org/10.1016/j.neuron.2010.03.003

Zhang, T., Gino, F., & Bazerman, M. H. (2014). Morality Rebooted: Exploring Simple Fixes to Our Moral Bugs. *SSRN Electronic Journal*, *34*, 63–79. http://doi.org/10.2139/ssrn.2427259

Appendix A. Items from the Civic Moral Disengagement Scale
(Caprara et al., 2009)

1. When there are no efficient refuse disposal services, there is no sense reproaching citizens who leave trash on the street
2. Some people are real disasters
3. To forget to declare a financial error in our favour is not serious, since it is the responsibility of the receiving person or institution to check for errors
4. There is no reason to fine those who draw ‘‘graffiti” on walls since others commit much more serious acts of vandalism
5. When traffic moves quickly, drivers who exceed the speed limit in order to keep up should not be fined
6. It doesn’t make sense for the individual to worry about environmental deterioration since the harmful effects are produced at the societal level
7. Evading taxes cannot be considered reprehensible considering the squandering of public money
8. Those who behave brutishly can only expect to be treated the same way by others
9. Thefts in large department stores are irrelevant compared to the stores’ earnings
10. Victims generally have trouble staying out of harm’s way
11. Thefts do not damage retail sales very much since insurance covers the losses
12. Drawing graffiti on walls is the expression of ‘‘creative spirit”
13. There is no sense feeling guilty for damages we have contributed to a problem if our contribution is a small part of the problem
14. Fraud in economic transactions is simply a ‘‘strategic distortion”
15. Silencing those who continue to be annoying, even using hard measures, is understandable
16. There is no sense in blaming individuals who evade a rule when everybody else does the same thing
17. Gambling is a passtime just like any other one
18. For the advance of science, it is lawful to use humans as ‘‘guinea pigs” even in high risk experiments
19. If people leave their belongings around, it is their fault if someone steals them
20. If someone loses control during a brawl, he/she is not completely responsible for the consequences of his/her actions
21. Citizens who litter the streets should not be severely persecuted since industry produces much more serious pollution
22. Using force is often inevitable to protect one’s own interests
23. Given the widespread corruption in society, one cannot disapprove of those who pay for favours
24. In order to keep family cohesion, its members should always be defended, even when they are guilty of serious crimes
25. Destroying old things is a way of convincing the state to provide new facilities
26. It is not the fault of drivers if they exceed the speed limit since cars are made to go at high speeds
27. Young people cannot be considered guilty if they smoke a joint since most adults use much stronger drugs
28. Rivals deserve being humiliated and maltreated
29. Loyalty involves not denouncing the transgressions committed by one’s friends
30. Employees are never responsible for executing the illegal decisions of their bosses
31. In order to force some people to work, they have to be treated like beasts of burden
32. Pornography is basically a cheap form of erotic activity

Appendix B. Items from the Moral Disengagement Scale

(Bandura et al., 1996)

1. It is alright to fight to protect your friends.
2. Slapping and shoving someone is just a way of joking.
3. Damaging some property is no big deal when you consider that others are beating people up.
4. A kid in a gang should not be blamed for the trouble the gang causes.
5. If kids are living under bad conditions they cannot be blamed for behaving aggressively.
6. It is okay to tell small lies because they don't really do any harm.
7. Some people deserve to be treated like animals.
8. If kids fight and misbehave in school it is their teacher's fault.
9. It is alright to beat someone who bad mouths your family.
10. To hit obnoxious classmates is just giving them "a lesson."
11. Stealing some money is not too serious compared to those who steal a lot of money.
12. A kid who only suggests breaking rules should not be blamed if other kids go ahead and do it.
13. Kids who are not taught to obey the rules should not be blamed for misbehaving.
14. Children do not mind being teased because it shows interest in them.
15. If someone acts like a jerk, it is okay to treat them badly.
16. If people are careless where they leave their things it is their own fault if they get stolen
17. It is alright to fight when your group's honour is threatened.
18. Taking someone's bicycle without their permission is just "borrowing it."
19. It is okay to insult a classmate because beating him/her is worse.
20. If a group decides together to do something harmful it is unfair to blame any kid in the group for it.
21. Kids cannot be blamed for using bad words when all their friends do it.
22. Teasing someone does not really hurt them.
23. Someone who is obnoxious does not deserve to be treated like a human being.
24. Kids who get mistreated usually do things that deserve it.
25. It is alright to lie to keep your friends out of trouble.
26. It is not a bad thing to "get high" once in a while.
27. Compared to the illegal things people do, taking some things from a store without paying for them is not very serious.
28. It is unfair to blame a child who had only a small part in the harm caused by a group.
29. Kids cannot be blamed for misbehaving if their friends pressured them to do it.
30. Insults among children do not hurt anyone.
31. Some people have to be treated roughly because they lack feelings that can be hurt.
32. Children are not at fault for misbehaving if their parents force them too much.

Appendix C. Items from the Social Desirability Scale

(Crowne & Marlowe, 1960)

Instructions: Listed below are a number of statements concerning personal attitudes and traits. Read each item and decide whether the statement is true or false as it pertains to you personally.:

If you have any questions please ask the researcher

1. Before voting I thoroughly investigate the qualifications of all the candidates.
2. I never hesitate to go out of my way to help someone in trouble.
3. It is sometimes hard for me to go on with my work if I am not encouraged.
4. I have never intensely disliked anyone.
5. On occasion I have had doubts about my ability to succeed in life.
6. I sometimes feel resentful when I don't get my way.
7. I am always careful about my manner of dress.
8. My table manners at home are as good as when I eat out in a restaurant.
9. If I could get into a movie without paying and be sure I was not seen I would probably do it.
10. On a few occasions, I have given up doing something because I thought too little of my ability.
11. I like to gossip at times.
12. There have been times when I felt like rebelling against people in authority even though I knew they were right.
13. No matter who I'm talking to, I'm always a good listener.
14. I can remember "playing sick" to get out of something.
15. There have been occasions when I took advantage of someone.
16. I'm always willing to admit it when I make a mistake.
17. I always try to practice what I preach.
18. I don't find it particularly difficult to get along with loud mouthed, obnoxious people.
19. I sometimes try to get even rather than forgive and forget.
20. When I don't know something I don't at all mind admitting it.
21. I am always courteous, even to people who are disagreeable.
22. At times I have really insisted on having things my own way.
23. There have been occasions when I felt like smashing things.
24. I would never think of letting someone else be punished for my wrongdoings.
25. I never resent being asked to return a favour.
26. I have never been irked when people expressed ideas very different from my own.
27. I never make a long trip without checking the safety of my car.
28. There have been times when I was quite jealous of the good fortune of others.
29. I have almost never felt the urge to tell someone off.
30. I am sometimes irritated by people who ask favours of me.
31. I have never felt that I was punished without cause.
32. I sometimes think when people have a misfortune they only got what they deserved.
33. I have never deliberately said something that hurt someone's feelings.

Appendix D. Items from the Levenson Psychopathy Scale

1. Success is based on survival of the fittest; I am not concerned about the losers.
2. I find myself in the same kinds of trouble, time after time.
3. For me, what's right is whatever I can get away with.
4. I am often bored.
5. In today's world, I feel justified in doing anything I can get away with to succeed.
6. I find that I am able to pursue one goal for a long time.
7. My main purpose in life is getting as many goodies as I can.
8. I don't plan anything very far in advance.
9. Making a lot of money is my most important goal.
10. I quickly lose interest in tasks I start.
11. I let others worry about higher values; my main concern is practical matters.
12. Most of my problems are due to the fact that other people just don't understand me.
13. People who are stupid enough to get ripped off usually deserve it.
14. Before I do anything, I carefully consider the possible consequences.
15. Looking out for myself is my top priority.
16. I have been in a lot of shouting matches with other people.
17. I tell other people what they want to hear so that they will do what I want them to do.
18. When I get frustrated, I often "let off steam" by blowing my top.
19. I would be upset if my success came at someone else's expense.
20. Love is overrated.
21. I often admire a really clever scam.
22. I make a point of trying not to hurt others in pursuit of my goals.
23. I enjoy manipulating other people's feelings.
24. I feel bad if my words or actions cause someone else to feel emotional pain.
25. Even if I were trying very hard to sell something, I wouldn't lie about it.
26. Cheating is not justified because it is unfair to others.

Appendix E. Standard Consent Form

I …… consent to participate in an experimental psychology study being run by Luis Manuel Silva supervised by Professor Dermot Barnes-Holmes in the Department of Psychology, National University of Ireland, Maynooth (Tel: +353 1 708 4765).

I understand and consent to the following:

o The experiment will not last longer than 2 hours.

o All data from the study will be treated confidentially. The data will be stored in a locked cabinet in the Department of Psychology and will be retained for a minimum of five years. An alphanumeric code will be entered into the IRAP program to protect my identity. This alphanumeric code will also be used on all explicit measures to protect my identity.

o Results from this research work will not be used deceptively or without my consent.

o My data is available to me at my discretion.

o I am free to terminate my participation in the study at any time and may withdraw the data obtained from my participation, if I so wish, up to the time of publication. If during my participation in the study I feel the information and guidelines I have been given are neglected or disregarded in anyway, or if I am unhappy about the process I may contact the Secretary of the National University of Ireland Maynooth Ethics Committee at pgdean@nuim.ie or 01 708 6018.

o I was given at least 24 hours before agreeing to volunteer for this study.

Please print and sign your name below if you are willing to abide fully by the conditions stated above.

Name: _____ Signature: _____ Date: _____

EXPERIMENTER:

I, Luis Manuel Silva, and Prof. Dermot Barnes-Holmes, as primary researchers, accept full responsibility for the care of all experimental participants and I confirm that all the necessary safety precautions have been taken.

Signature of experimenter: _____ Date: _____

Appendix F. Disclosure Information Sheet

This study. In the study you just took part in, we measure attitudes and beliefs towards certain social situations and opinions about "moral" situations – that is, situations in which people can cause benefit or harm to others. Beyond math skills, what we intend with the mental math task is to check whether participants deliberately wait to see the answer at least once.

Rationale. If we had told you that this is what we were going to measure, you would have probably given it a second thought, and avoided doing it even if willing to, but since we are interested in what people ordinarily do, we wanted to create a situation in which you could feel comfortable to decide whether you wanted to wait for the answer or not. Studies involving mild deception, such as this one, are permitted in cases where full disclosure may affect the performance on a critical task, and provided that said full disclosure is performed after the completion of the tasks (which is why we give you this sheet).

Your feelings. There is nothing inherently wrong in having decided to wait for the answer, nor inherently right in not doing it, so you should not feel that your choice implies that you are a fair player or not. This is not a psychological test – it is not, and it actually cannot, be used to make general conclusions regarding your personality. Studies carried in the Department of Psychology are fully compliant with several Codes of Professional Ethics, and you can be sure that you are not being judged in any way for your responses.

Your data. Personal information that may enable someone to identify you has not been collected. The computer never asked for your name, address (postal or electronic), course of study, or other sensible information – we only have your responses, your age and your gender. This guarantees that your responses can never be matched to your name. For this reason, we are unable to provide feedback regarding your own performance – the analysis will be done with a group of participants, not on a case-by-case basis.

Publishing. The results of this study will be part of the doctorate thesis of the researcher, and some data will probably appear in a paper in a peer-reviewed scientific journal. It should be emphasised that, since no personal information was collected, there is no way you can be identified.

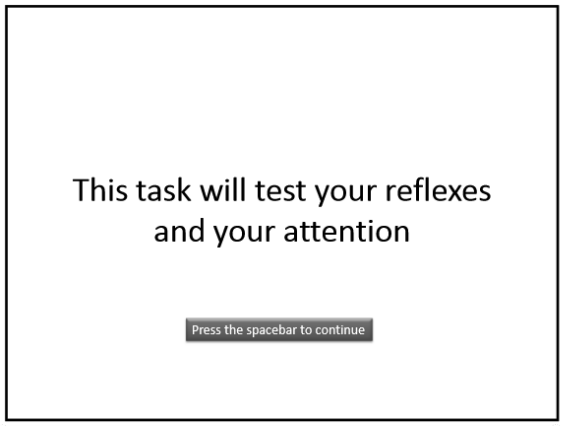
Your rights. Despite the protection of your data, if you decide now that you would not like to have your responses recorded and analysed you can have your responses erased. If this is your wish, the researcher will show you to a paper shredder where you can destroy the response sheets, and he will delete your data from the computer in your presence. However, before making the decision, please remember, once again, that your responses cannot be matched to your name in any way. You will not forfeit your incentive by having your data erased. It will also not affect your subsequent participation in other studies should you want to volunteer for them in the future.

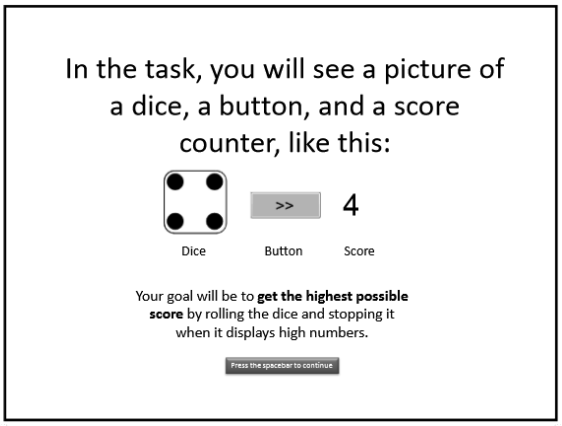
If you have any questions whatsoever at this point about the study, please feel free to ask the researcher.

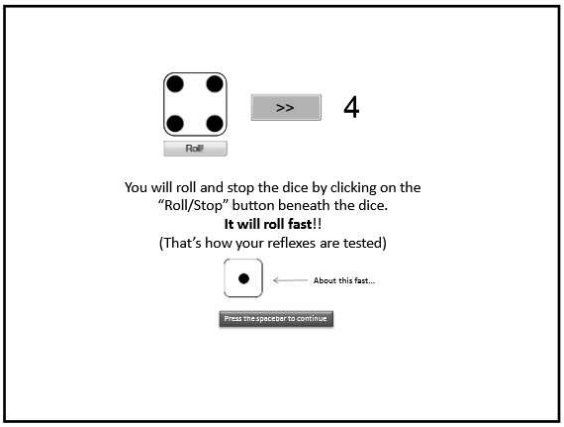
I understand the true nature of the study and the reason for its temporary concealment, and agree with it. I understand how the data will be handled and I agree with it.

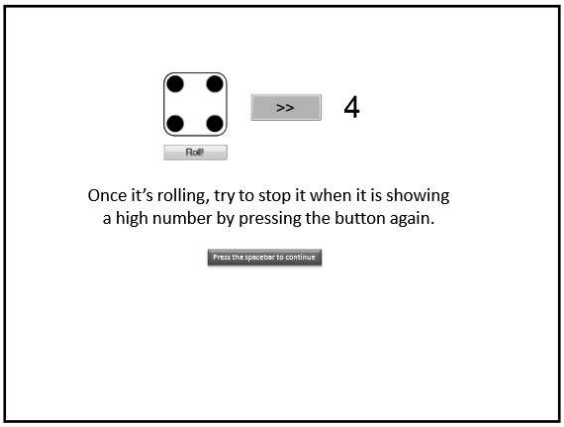
Name: _____ Date: _____ Signature: _____

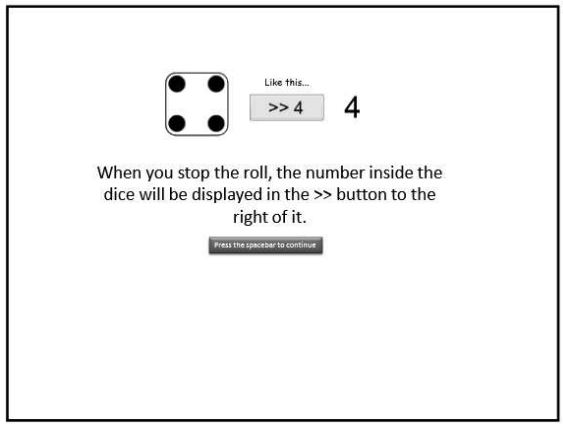
Appendix G. Training Slides for the Dice Cheating Task

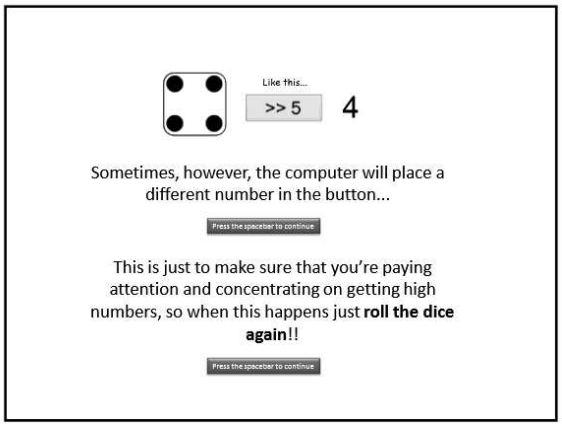












<click!>

>> 3

3

For example, in this case you've stopped the roll when it was showing a 3 and the >> button also shows 3.

In this case, you would click the >> button, and that would add 3 to your score (the button then goes red to indicate that you've already clicked it).

Press the Spacebar to continue

>> 3

This case, however, is a *non-match*: the roll stopped when it was showing five, but the button shows three.

As said before, the computer will sometimes do this to verify that you are paying attention. Remember, when this happens just press the Spacebar or click the "Roll/Stop" button (i.e., roll the dice).

Next

The task will finish when the maximum possible score has been reached (400 – really difficult!) or after a fixed number of trials. At the end, you will be asked a couple of questions about the task.

Press the Spacebar to continue

So, here are the main points again:

1. The main goal is to **get a high score**.
2. So try to stop the dice when it shows **high numbers**.
3. Add points only when there is a **match** between dice and button.

Press the spacebar to continue

The task is about to start. Please make sure that you understand this instructions. If you have ANY questions call the experimenter so he can answer them.

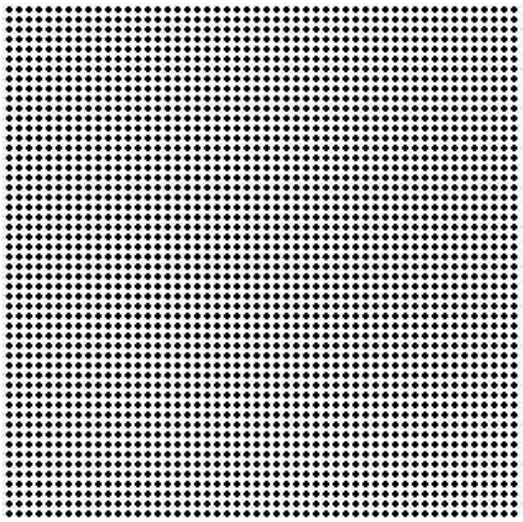
The first few trials will be just practice, so the computer will give you a warning if you add scores when the numbers on the dice and the button do not match.

The computer will let you know when practice mode has ended. When it does, you will no longer get any feedback, so pay close attention!!

Press the spacebar to begin

Appendix H. Mortality Salience Induction Procedure

Sample figure containing number of dots (weeks left to live) for a female participant aged 18:



Script for the intervention (experimental group)

(DOTS SECTION BEGINS)

I think it's often very easy to forget just how short life is, especially for young healthy students.

To help convey this, I've put together this diagram for you.

Given that I know your age and gender, it's trivial for me to estimate your expected lifespan. [place sheet in front of them]

Based on that, the number of dots on this piece of paper is equal to the number of weeks you have left to live. [said very slowly and carefully, and then a long pause]

I promise that I'm not trying to trick or deceive you. It's a surprisingly small number of dots, isn't it?

The thing about dots is that once you spend them you can't get them back. This is not a rehearsal, you will not get a second shot. This is your life, right now, ending, one day at a time.

The other thing about dots is that they run out, no matter what you do. Make no mistake, death is coming.

You have a limited number of days left on this planet, and, like all of us, you're faced with the difficult question of what you're going to do with them.

How many of these dots will be well spent dots, doing things that you truly value, like time with friends and family, and how many dots will be more like hovering dots and X-factor dots?

(QUESTIONS SECTION BEGINS)

With all of that in mind, I'd like you to write out a few lines about what you think dying itself will be like.

[Give them sheet containing the following questions:

1. What emotions does the thought of your own death arouse in you?

2. Jot down, as specifically as you can, what you think will happen to you physically as you die and once you are physically dead.

3. "The one thing I fear most about my death is…"

4. "My scariest thoughts about death are…"]

Post experiment debrief:

Discuss with participant how, although death is inevitable, it has also been said that it is the ultimate motivator in life, or even "the mirror in which meaning in life is reflected". By knowing that we have a limited number of dots to spend, we are motivated to spend them in ways we value.

Script for control group

I think it's often very easy to forget just how big our solar system is, especially for young students.

To help convey this, I've put together this diagram for you. [place sheet in front of them]

The number of dots on this piece of paper is equal to the number of million kilometres that separates the Sun and the Earth. [said very slowly and carefully, and then a long pause]

I promise that I'm not trying to trick or deceive you. It's a surprising number of dots, isn't it?

I'd like you to write out a few lines about what you think about the size of our solar system.

1. What emotions does the thought of the size of our solar system arouse in you?

2. Jot down, as specifically as you can, whether you think mankind will be able to travel throughout our solar system.

3. "The one thing that comes to mind when pondering the size of our solar system is…"

4. “My thoughts about our solar system are…”

Appendix I. Spanish Translations of the MSI used in Study 6

(SECCIÓN DE PUNTOS)

Creo que suele ser muy fácil olvidar lo corta que es la vida, especialmente para estudiantes jóvenes y sanos.

Para ayudarte a comprender esto, he creado este diagrama.

Dado que conozco tu edad y tu sexo, puedo estimar fácilmente tu expectativa de vida [poner hoja frente al participante]

Con base en eso, el número de puntos en esta hoja de papel representa *el número de semanas que te quedan de vida* [decirlo lenta y cuidadosamente, y luego una pausa larga).

Te aseguro que no estoy tratando de engañarte con esto. Es un número sorprendentemente pequeño de puntos, ¿no es verdad?

La cosa con estos puntos es que una vez que los gastas no los puedes tener de nuevo. Esto no es un ensayo, no habrá una segunda oportunidad. Esta es tu vida, ahora mismo, acabándose día a día.

Otra cosa de estos puntos es que van a acabar, sin importar lo que hagas. La muerte llegará – no pienses que no.

¿Cuántos de estos puntos vas a gastar bien, haciendo cosas que de verdad valoras, como pasar tiempo con tu familia, y cuántos serán puntos haciendo pereza y viendo “Yo me llamo”?

(SECCIÓN DE PREGUNTAS)

Me gustaría que escribieras algunas líneas sobre cómo piensas que será morir algún día.

[Darles la hoja de las preguntas]:

1. ¿Qué emociones te genera pensar sobre tu propia muerte?
2. Escribe, con tanto detalle como puedas, qué crees que te pasará físicamente cuando mueras y cuando estés físicamente muerto(a)
3. “Lo que más me asusta de mi muerte es…”
4. “Mis pensamientos más aterrorizantes sobre la muerte son…”